Analysis the Likable Characteristic by Marvel

Rebecca Li
Student ID: A58371398

Final Report

Figure 1: Marvel Comics

1 Introduction

Marvel is a famous comic that created by Stan Lee and this comic has been filmed to lot's of movies, the story in Marvel Universe and the thousands of characters brings happiness to people around the world. This project aims to use method of lm, lda, qda analysis the the likable and unlikable characteristics of character in Marvel, find the pattern in the characteristic and explore the reason of those character being created. Characteristics like eye color, hair color and sex…etc will be use to predict the alignment of the character and number of appearance, time of appearance of character may also relate to this.

2 Related Work

Marvel has thousands of characters and each character has different outlook and own characteristics, the main character that has most number of appearance is Iron Man and it's also the most popular character in Marvel, by comparing with the look of Iron Man and author Stan Lee, their are some similarities.

Figure 2: Iron Man

Figure 3: Stan Lee

I don't find research about why Stan Lee create the main character has some similarities like himself or this similarity is coincidence. However, in StackExchange, there is a discussion raised about "Do authors often base their characters off of themselves.". From this discussion, there is some example like

"Temperance Brennan" in the *Bones,* the narrator in *Justine,* "Hermione Granger" in the *Harry Potter and* "Esther Greenwood" in *The Bell Jar,* which shows that it's not raw main and especially good character who has aspects of author. This means the physical and mental characteristics are meaningful.

There is a research about the eye color frequency of Marvel superhero, from this research, blue and brown are the most frequently eye color in Marvel, this may not enough to support my prediction about similarity between author and main characters. Because those two eye colors are most frequently appear in human. So I will improve this by also analysis other characteristics like hair color.

In order to let reader appear the motion with characters and make the story attractive, what makes a character good or bad is important. In general, readers are tend to like good character and hate bad character. Article *The Liking-Similarity Effect: Perceptions of Similarity as a Function of Liking* shows there is a pattern that people tend to like the person who has similarity with themselves.

3 Dataset

(i) The data used in this project include information of most characters appeared in marvel, it include 16376 characters and 13 columns include name, page id, hair color, eye color…etc.

(ii) (ii) After delete incomplete row, there are 4353 characters are valid data. I picked variables related to likable like eye color, hair color, sex, and number of appearance in comic to predict the align of the character.

(iii) Since those variables are factors, and include many types, I combined the minority types to one type called "Other" in order to keep the balance of number of each type and avoid bias. For variable sex, since there are only few of sex other than Female and Male, if set those to Other, would influence the accuracy of model, I decide to only count female and male characters

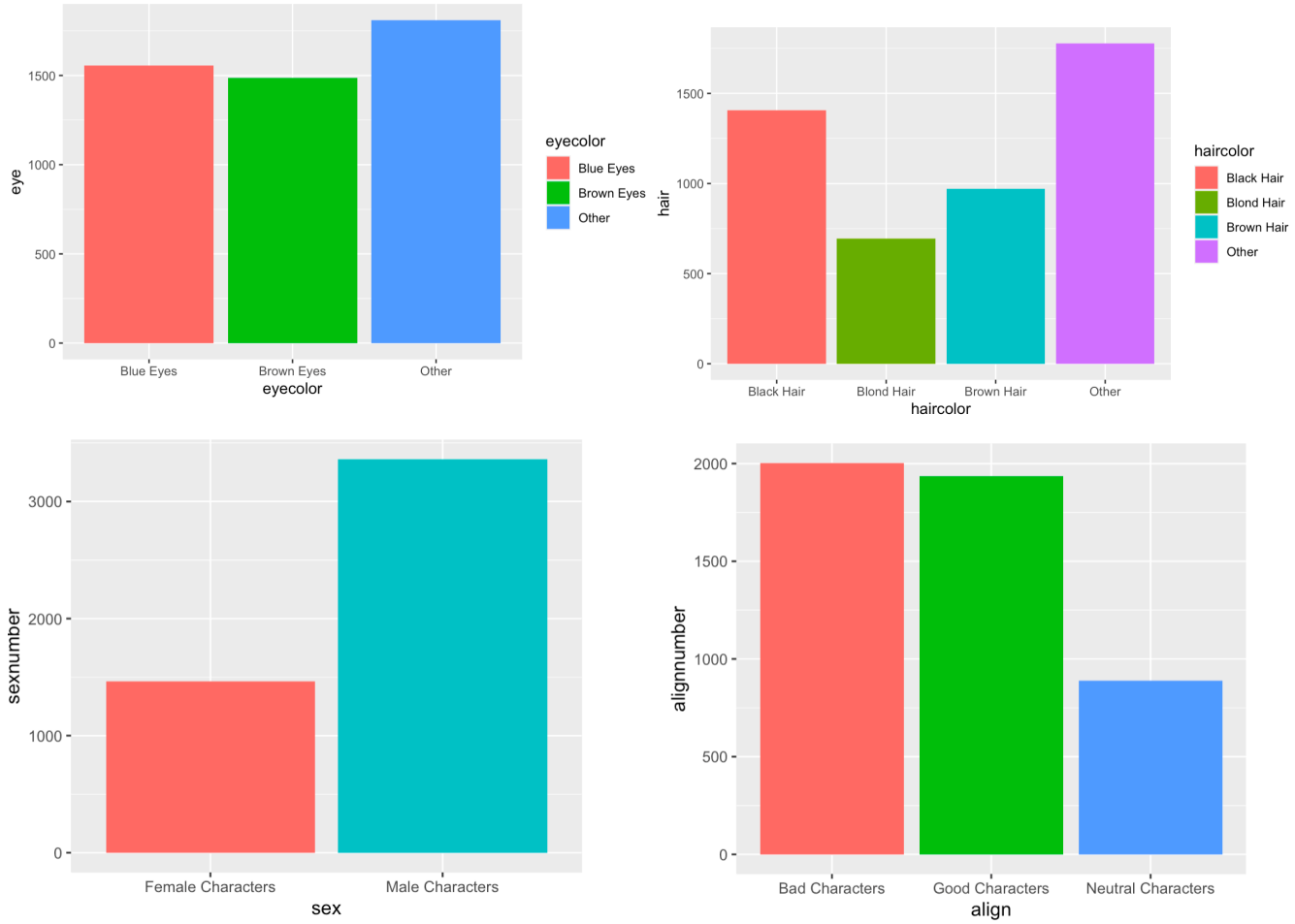

Figure 4: Eye color, Hair color, Sex, Align

4 Method

This project used 3 regression model to train the data and make prediction of Align, find the significant factor that influence the Align and find the model that relatively better fitted data by comparing the accuracy of result from linear regression, LDA, and QDA, (i) Linear Regression: find the significant level of variable and adjust the variable used for prediction. (ii) LDA and QDA, method to train the model by using the separated train data, the ratio of train and test data are 2:1, predict the Align by those method and show the table of accuracy, comparing those two method. Detail information are provided in *Final project code.Rmd.*

5 Results

Linear Regression:

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.0875 -0.4152 -0.0132  0.4561  0.6769

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.150e-01  1.467e-02  48.725   <2e-16 ***
EYE         -1.809e-01  1.526e-02 -11.854   <2e-16 ***
SEX         -1.734e-01  1.369e-02 -12.670   <2e-16 ***
HAIR        -3.797e-02  1.492e-02  -2.545   0.0109 *
APPEARANCES  3.417e-04  3.767e-05   9.071   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.432 on 4822 degrees of freedom
Multiple R-squared:  0.08573,   Adjusted R-squared:  0.08497
F-statistic:   113 on 4 and 4822 DF,  p-value: < 2.2e-16
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.0615 -0.4297 -0.0231  0.4650  0.6643

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.010e-01  1.361e-02  51.497   <2e-16 ***
EYE         -1.874e-01  1.505e-02 -12.455   <2e-16 ***
SEX         -1.782e-01  1.356e-02 -13.143   <2e-16 ***
APPEARANCES  3.393e-04  3.768e-05   9.006   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4323 on 4823 degrees of freedom
Multiple R-squared:  0.0845,    Adjusted R-squared:  0.08393
F-statistic: 148.4 on 3 and 4823 DF,  p-value: < 2.2e-16
```

Based on the result, p-value of each factor are really low but R-squared value is pretty low, which means that all of the variables are significant but the model does not fit the data, in order to prove it, I take the Hair Color variable away because it has the relatively largest p-value, which means it is less significant to Align compare with others. After remove the Hair Color variable, the fitness level of model does not increase, which means leaner regression model is not suitable for this data.

LDA $ QDA:

```
test_y   0 0.5   1
   0   444   0 240
   0.5 171   0 202
   1   222   0 331
[1] 0.4813665
```

```
test_y   0 0.5   1
   0   684   0   0
   0.5 373   0   0
   1   553   0   0
[1] 0.4248447
```

Based on the table of computation between prediction from model and data, LDA has higher accuracy, however, from the table, there is no prediction result in 0.5 which is Neutral Character, which is quit understandable because the number of character belongs to Neutral is much less than the characters belongs to other Align. QDA only has prediction on 0, which means the model is bias and has low reliability and need to improve even it has similar accuracy with LDA.

6 Discussion and Conclusion

Based on the result, LDA has the best accuracy compare with other model, I think there are some thing could be improve in the model, the result of model does not high level of accuracy, it may because the variables are not enough to make prediction even they are all significant. Also it's could improve on the way of setting the factors, I set categories from same variables from 0 to 1, like Align, good represented by 1, Neutral represented by 0.5 and Bad represented by 0. The small variation between the value of factor may lead to the low correlation between p-value with the real significance. However, I think the project does proved that there is strong relation between characteristic and the likability.

7 Reference

[1] Figure 1:
https://www.google.com/url?sa=i&url=https%3A%2F%2Fm.comixology.com%2FOrigins-Of-Marvel-Comics%2Fdigital-comic%2F697755&psig=AOvVaw3kigCRlvQNTQsPP5--ODJt&ust=1618828667789000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCLirhvTMh_ACFQAAAAAdAAAAABAP

[2] Figure 2:
https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.quora.com%2FHow-many-Iron-Man-suits-are-there-in-the-comics&psig=AOvVaw0AJcPG_cyv86V5N7eKCsxF&ust=1618833140569000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCKjm7cfdh_ACFQAAAAAdAAAAABAi

[3] Figure3:
https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.plymouth.edu%2Ftheclock%2Fstan-lee-a-marvel-of-our-time%2F&psig=AOvVaw2YzJ5kanUCYF2f5KaKb2GD&ust=1618833546622000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCPiQk4nfh_ACFQAAAAAdAAAAABAD

[4] "Do authors often base their characters off of themselves.", StackExchange:
https://writing.stackexchange.com/questions/41217/do-authors-often-base-their-characters-off-of-themselves

[5] Eye color Frequency
https://www.kaggle.com/habit456/marvel-superhero-eye-color-frequency/data?select=marvel-wikia-data.csv

[6] *The Liking-Similarity Effect: Perceptions of Similarity as a Function of Liking.* Brian Collisson, Jennifer L. Howell
https://www.tandfonline.com/doi/abs/10.1080/00224545.2014.914882

[7] Marvel Dataset
https://www.kaggle.com/fivethirtyeight/fivethirtyeight-comic-characters-dataset