R Notebook

Contents

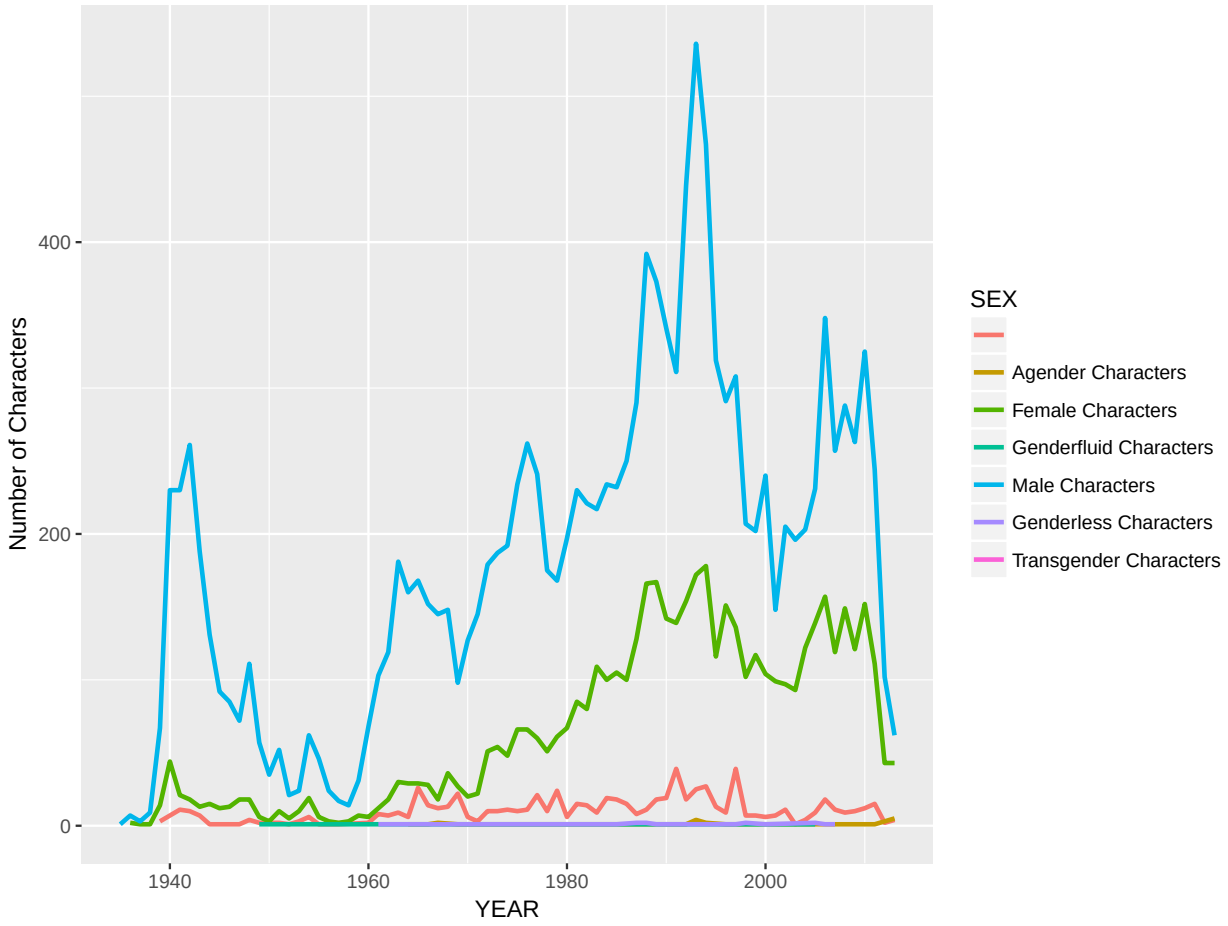
```
options(repos="https://cran.rstudio.com" )
```

This dataset describes all the DC and Marvel comic book characters that have appeared till 2013. Originally available on DC and Marvel Wikias, they were scrapped for similar analysis purposes by the website fivethrityeight.com.

The plot below gives the number of characters, belonging to different sexes, that were introduced each year since 1935.

For generating this plot I needed the characters to be grouped by their SEX and the Year in which they were introduced. I have used the 'Aggregate' function for this purpose. For calculating the sum of the characters introduced in any year I have applied the function length() on the 'Comics' field.

```
ggplot(Gender, aes(YEAR, Comics, col = SEX)) + geom_freqpoly(stat = "identity", lwd = 1) + ylab("Number
```




```
#ggplotly()
```

Clearly, and as expected, there is ratio of males to females is heavily skewed with the males highly outnumbering the females.

0.0.1 Sex ratio of the genders

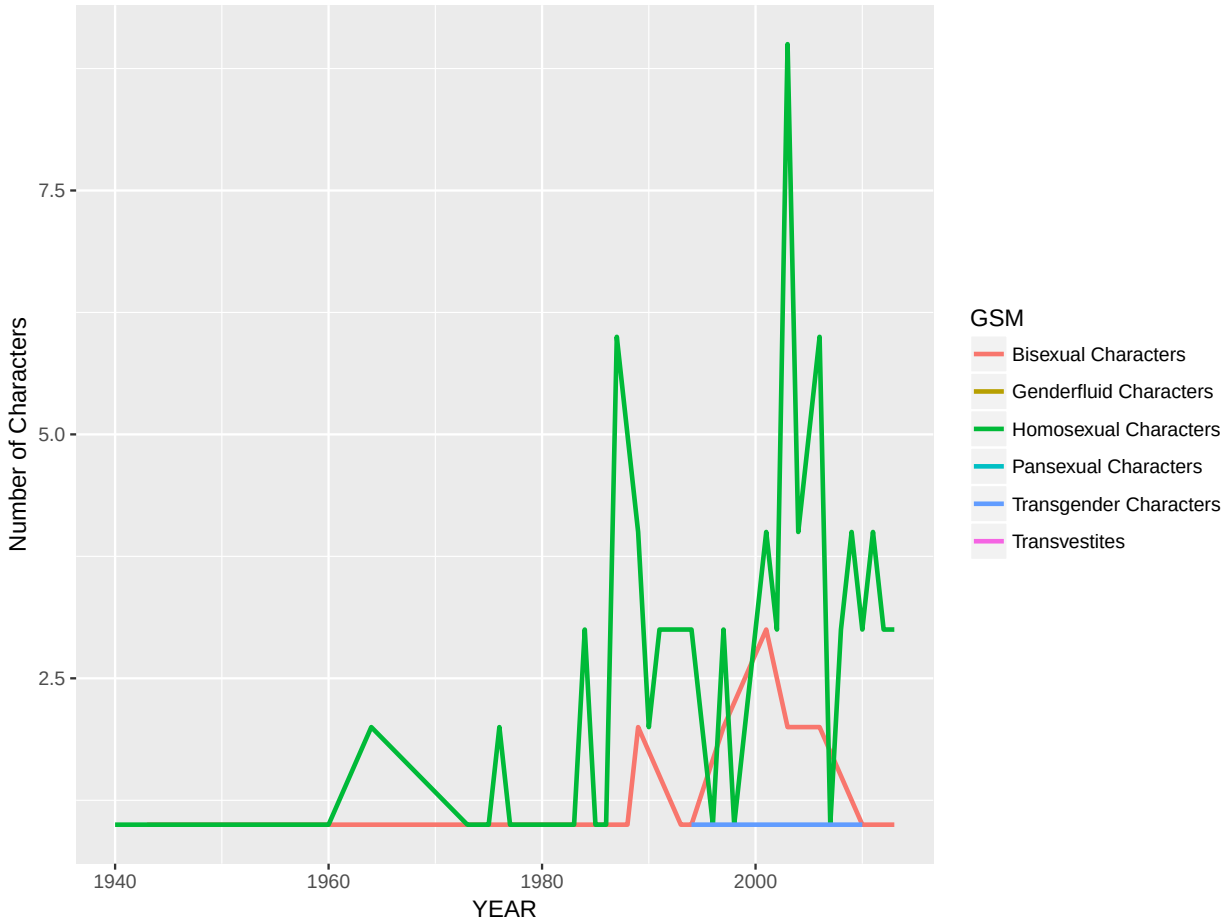
0.0.2 Sexual Preference of Characters

Below is a plot that describes the sexual preferences of the characters introduced each year since 1935.

For this purpose, I've again used the aggregate function with the Function 'length' applied on Comics to give the sum of the count of the characters introduced in any year.

```
Inclination <- aggregate(Comics ~ GSM + Comics + YEAR, data = FullComics, FUN = length)
Inclination$GSM[Inclination$GSM == ''] <- NA
Inclination <- na.omit(Inclination)
#table(Inclination$GSM)
```

```
ggplot(Inclination, aes(YEAR, Comics, col = GSM)) + geom_freqpoly(stat = "identity", lwd = 1) + ylab("Nu
```




```
#ggplotly()
```

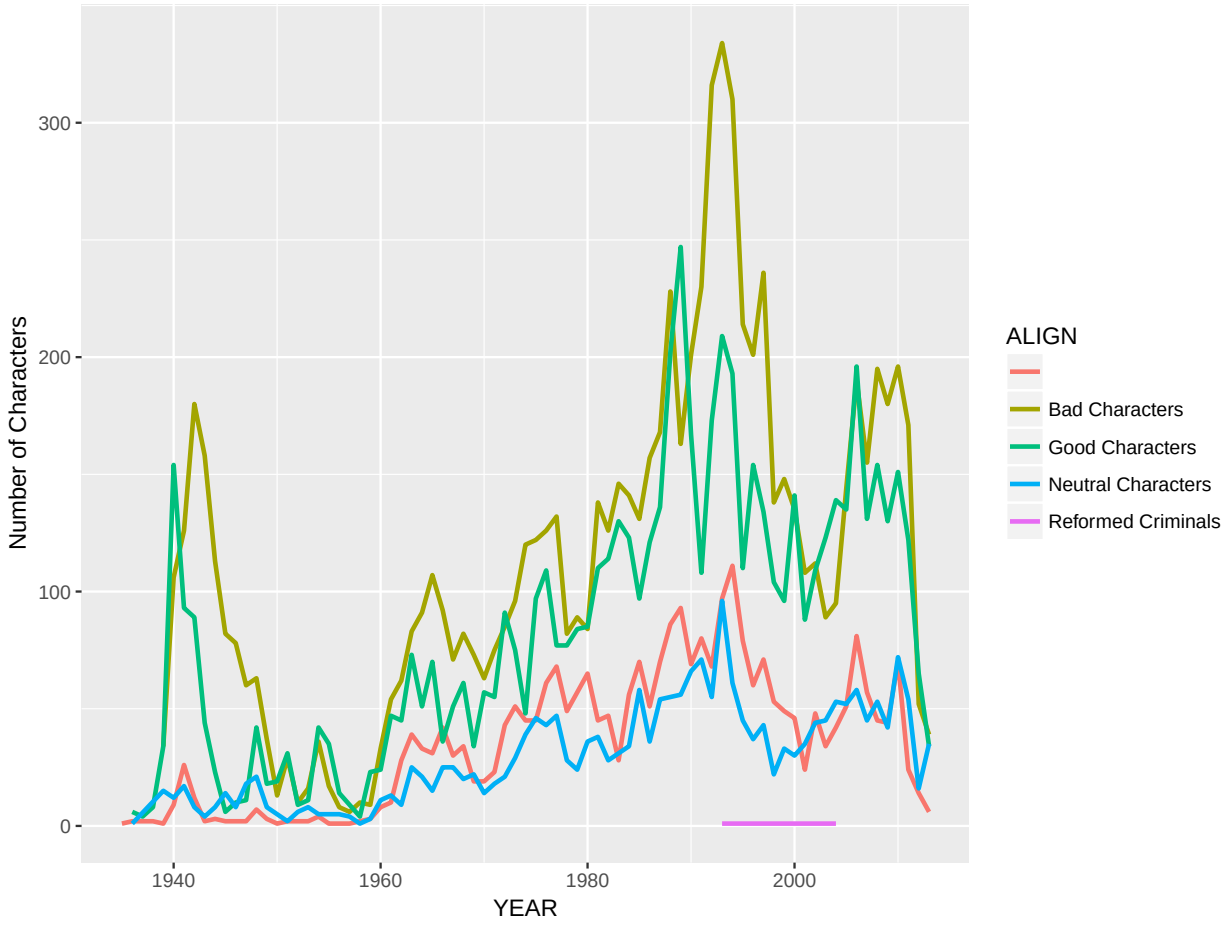
It was expected that the creators of the comic series would become increasingly more open towards introducing non-hetrosexual characters with time. As expected, we see a spike in such characters since 1990. On performing a search about the events that took place in the 1990's following information was found:

1990 : Source Wikipedia Decriminalisation of homosexuality: UK Crown Dependency of Jersey and the Australian state of Queensland LGBT Organizations founded: BiNet USA (USA), OutRage! (UK) and Queer Nation (USA) Homosexuality no longer an illness: The World Health Organization

These reasons, I believe, could be the reasons for the spike in the introduction of non-hetro-sexual characters.

The following plot gives the number of bad Vs good characters introduced over time.

```
good <- aggregate(Comics ~ ALIGN + YEAR, data = FullComics, FUN = length)
#good
ggplot(good, aes(YEAR, Comics, col = ALIGN)) + geom_freqpoly(stat = "identity", lwd = 1) + ylab("Number
```



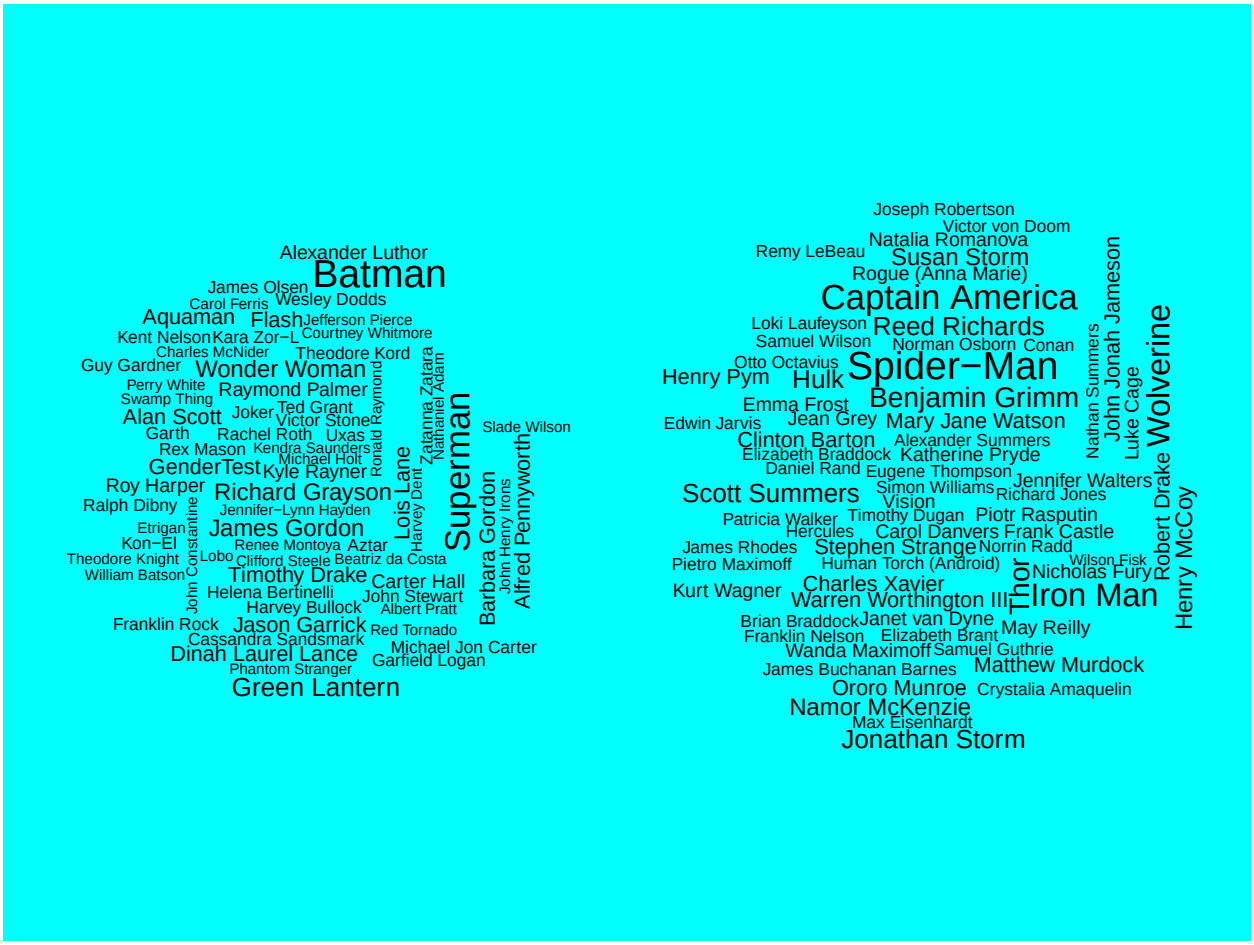

```
#ggplotly()
```

As expected, the number of Bad characters always exceeds the number of Good characters, especially, during the years 1990-1995.

0.0.3 Wordcloud - representation of the most promminent characters from both the Universes.

Following is a WordCloud of the most important characters in both the universes. This representation suits better than a histogram since it can incorporate far many characters in a much smaller space.

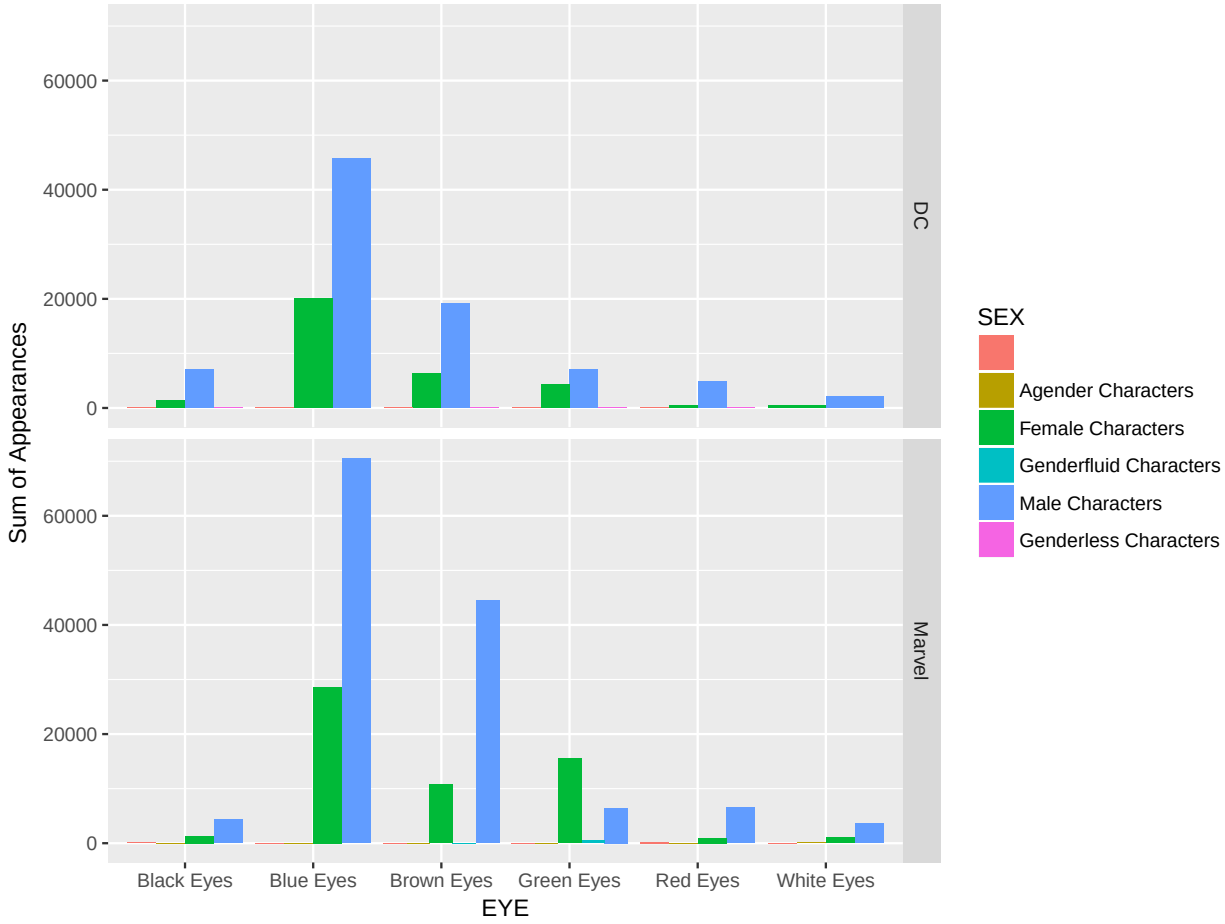
```
op <- par(mar=c(1, 2, 2, 1),mfrow=c(1, 2),bg="cyan")
wordcloud(DC_comics$Name_Commplete,DC_comics$APPEARANCES, max.words = 70, random.color = FALSE, scale =
wordcloud(Marvel_comics$Name_Commplete, Marvel_comics$APPEARANCES, max.words = 70, scale = c(1.5, 0.5))
```



As evident from the WordCloud, the DC Universe rides heavily on the shoulders of Batman and Superman characters followed by Green Lantern and Wonder Woman. Whereas, in case of Marvel Spider-Man stands out from the rest, followed by Iron Man and others.

The following plots highlight the most desirable eye and hair color in any comic book character. For this plot, I have taken only the five most prominent EYE and HAIR colors prevalent in the characters. I've used the subset function for extracting only those characters that had the EYES and HAIR colors.

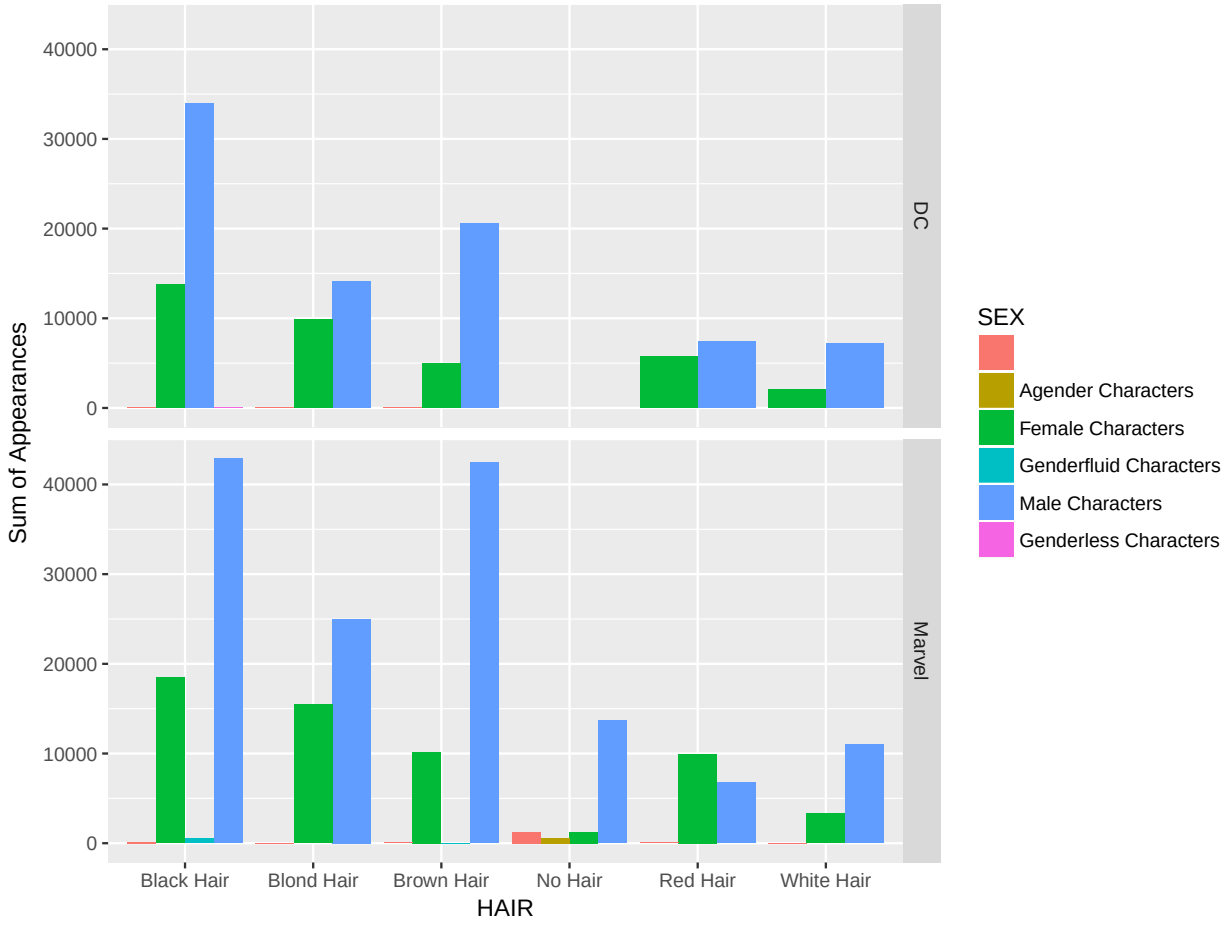
```
ggplot(eyes, aes(EYE,APPEARANCES, fill = SEX)) + geom_bar(stat = "identity", position = "dodge") + ylab
```




```
#ggplotly()
```

For males, in decreasing order of appeal, the eye colors are - Blue, Brown, Black, Red and White For Females they are -Blue, Green, Brown, Black, White and Red. This distribution is same for both the production houses.

```
ggplot(hair, aes(HAIR,APPEARANCES, fill = SEX)) + geom_bar(stat = "identity", position = "dodge") + ylab
```



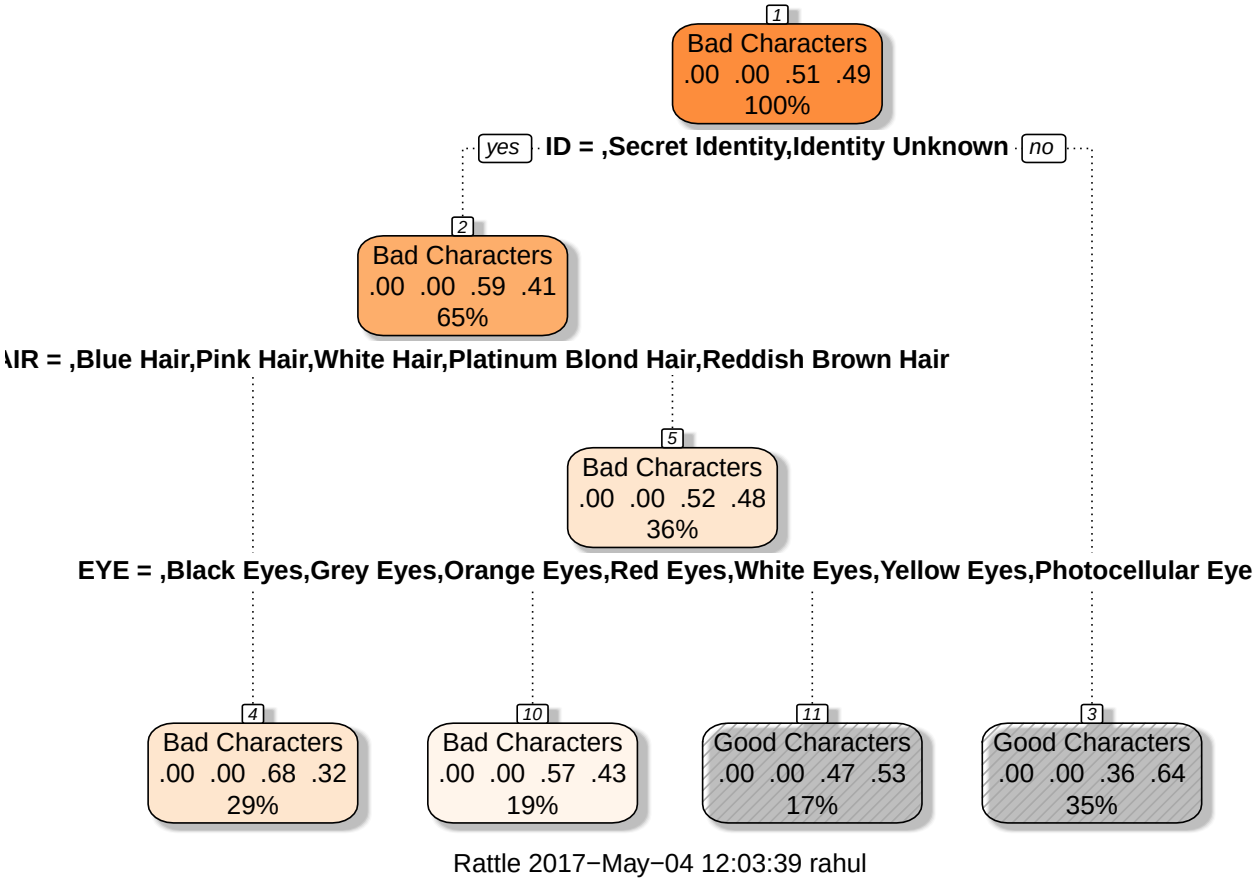

```
#ggplotly()
```

For males, in decreasing order of appeal, the hair colors are - Black, Brown, Blond, White, Red, No Hair. For Females they are -Black, Blond, Red, Brown and White with a very few without hair.

0.0.3.1 Predicting character alignment based on facial features and identity.

Following is an analysis of how prominent are certain facial features (eye and hair color) and the type of identity (secret or public) in determining whether a character is good or bad. For this analysis the dataset was segregated into the two different production houses - Marvel and DC so that a comparitive analysis could be done. A decision tree analysis was done on R after which the decision tree was graphically plotted.

```
##                                B     Bad Characters
##             2803               1               8759
##    Good Characters Neutral Characters Reformed Criminals
##             6690               2342                  3
```




```
## [1] 0.6165939
```

Interpreting this tree diagram : 1>. The title of each node tells us the type of character that has the highest probability of being found in the set given that the conditions stated just above the node are true. For example - in the case of the first node there is no condition stated prior to the first node. The first node is titled "Bad Characters" indicating that on the whole bad characters outnumber the good characters.

2>. The numbers on the line next to the title of each node give the relative probability of all the different type of characters in that subset, with the highest number corresponding to the type of character mentioned as the title of the node. For example in the first node the "Bad Characters" have a probability of 0.50 which is equal to that of the "Good Characters" - also 0.50. In the case of the first node the probability of a character being Bad exceeds that of being Good only by a small fraction - which is lost while rounding off.

3>. The number expressed as a percent on the line next to the set of numbers representing the probabilities gives the percent of the data-set that is explained by that node. For example the first node has this number as 100% indicating that it represents all the characters of the data-set.

4>. The statement written next to each node indicates the condition on which the next split is made. It has two outcomes - "Yes" or a "No"- Based on which the next node is selected for analysis. For example - for the first node the statement is "ID = Secret Identity, Identity Unknown". If this condition is TRUE i.e "Yes" one goes to the node on the left - "Bad Characters", if not, i.e. "No" one goes to the right node - "Good Characters".

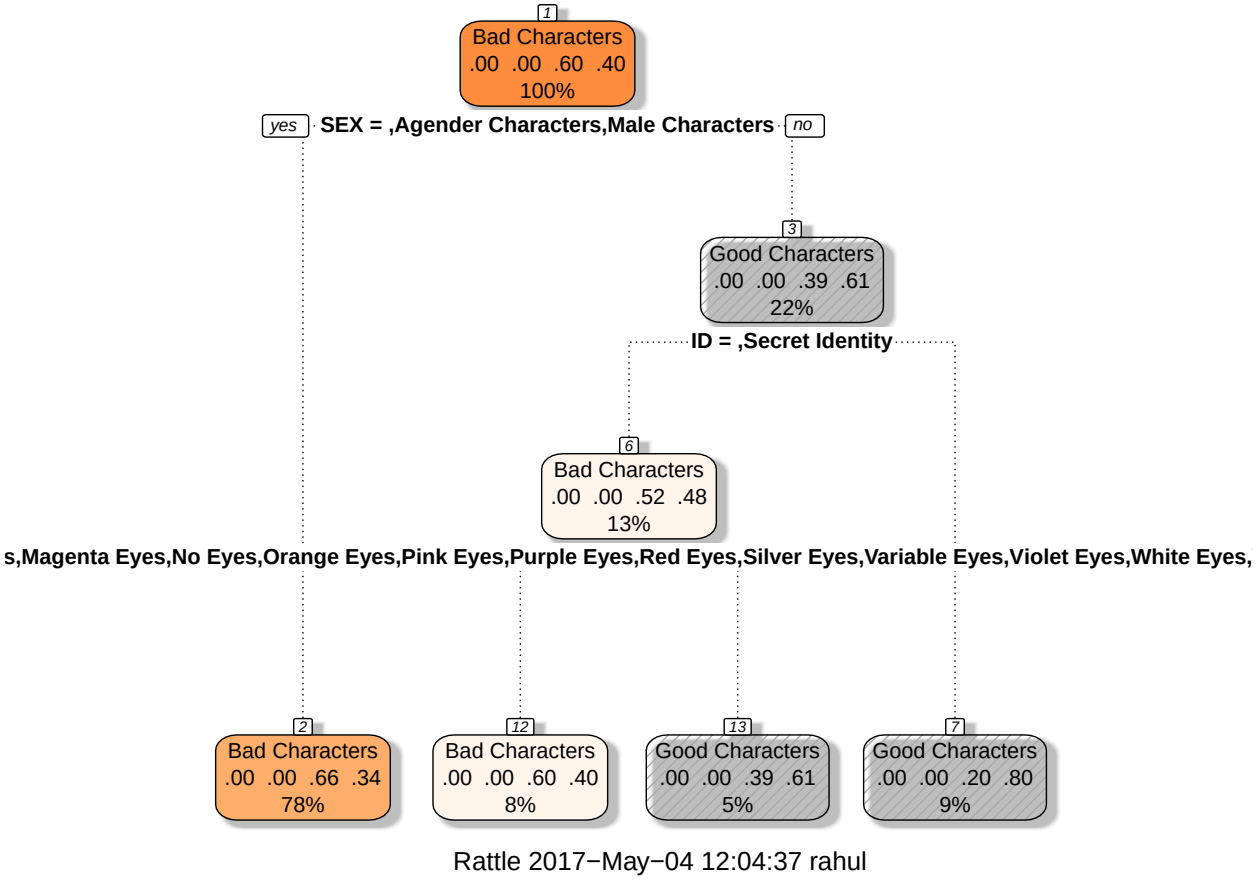
Conclusions from this tree diagram:

1>. Node.1. The "Bad Characters" slightly outnumber the "Good Characters" - both having a probability of 50-50. 2>. Node.2. If only those characters having a "Secret Identity" or an identity which is unknown are

selected - a total of 65% of the characters come under this group out of which there is a high probability 0.59 of the characters being "Bad Characters". 3>. Node.3. Characters whose identities are not secret or unknown comprise of 35% of the total and in them there is a high probability that any specific character is good (0.65). 4>. Node.4. If only those characters are selected which are having either Black, Grey, Hazel,Orange, Red, White such characters comprise a total of 44% of the total characters and out of them a strong possibility is that such characters are bad(0.64) as compared to good (0.36). 5>. Node.5. Characters which do not have the above mentioned eye colors, i.e. thoose having Blue, Brown or Green eyes constitute a total of 20% of all DC characters and there is a 53% probability that the character is "Good".

Takeaways from this analysis: 1>. As expected, the Bad guys generally have a secret or an unknown identity whereas the good guys mostly don't. 2>.Except for the "Black" eye color most of the other eye colors are very rare if not impossible to find naturally. This, probably, has been done to give an element of exclusivity or other-worldly appearance to the Bad Characters.

A fact to be noted is that the accuracy of this model is around 60% when only Hair color, Eye color and Sex is taken into account and around 61% when "Identity" is also added to the model. This indicates that "Identity" is not a major factor when deciding which character is bad or good when it comes to DC comics.




```
## [1] 0.6162551
```