Are Comics Becoming More Diverse?

By Conor McGuire

Problem Statement

I want to investigate whether Comics are becoming more diverse.
With the recent surge of popularity of superhero movies, more people are starting to read comics, but are comics representing everyone fairly?

Sources And Metadata

Sources

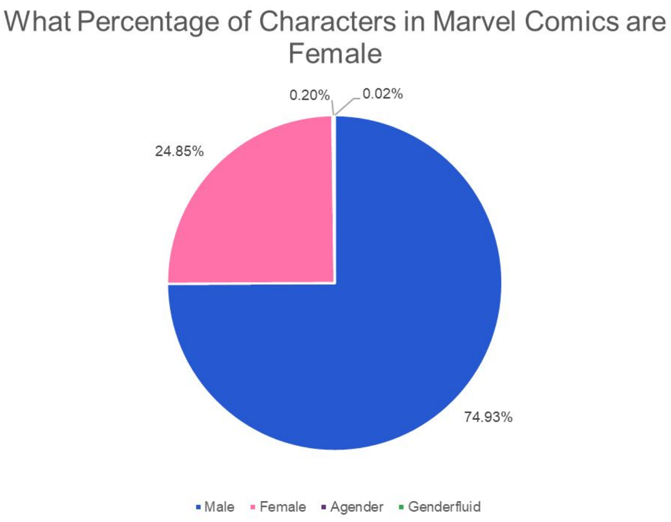
For this project I decided to analyse the dataset entitled “marvel-wikia-data” which can be found at the following link: https://github.com/fivethirtyeight/data/tree/master/comic-characters.

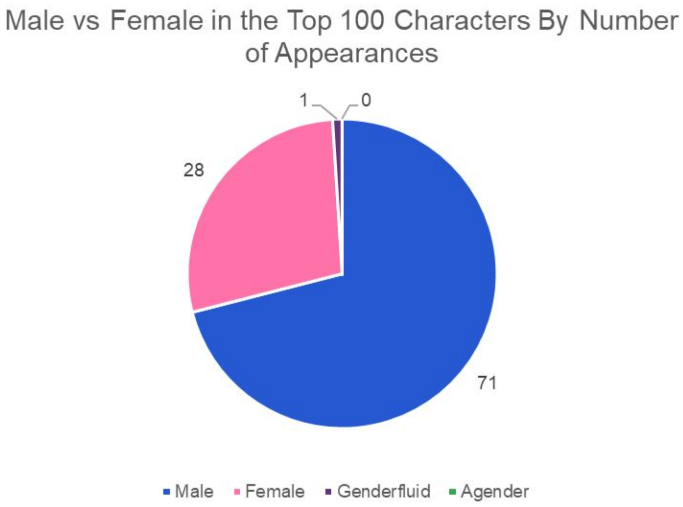
I chose this dataset because I have an interest in the Marvel movies that are based on these comic books.

Metadata

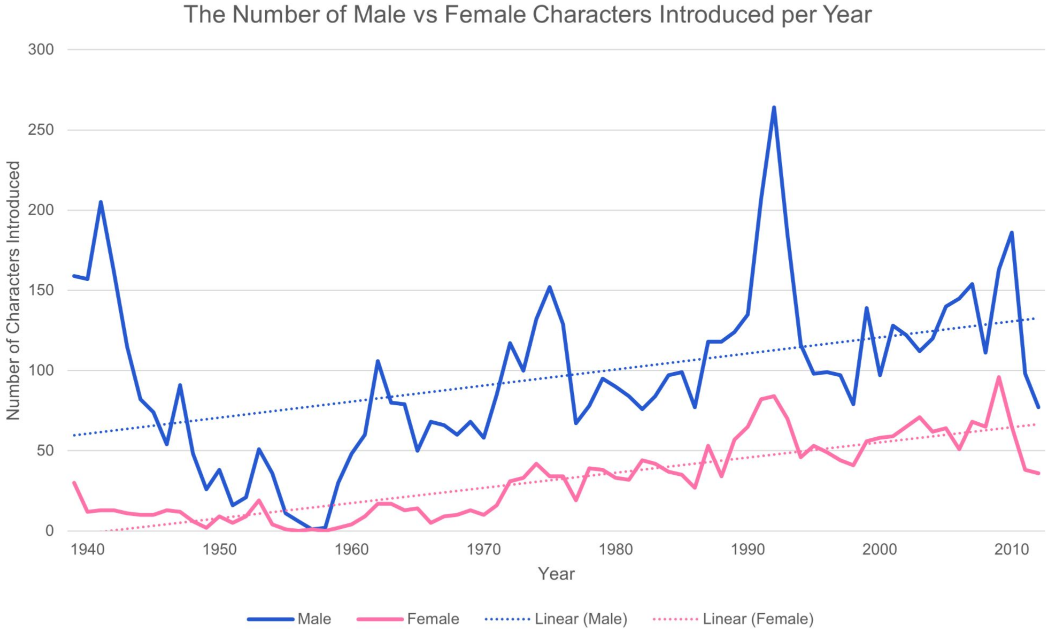
Title: marvel-wikia-data
Description: data behind the story “Comic Books Are Still Made By Men, For Men And About Men.” Scraped on August 24th 2014
Author: Andrew Flowers
Number of columns: 13
Number of rows: 16,376
Type of data: Includes Categorical Text Data, Numerical Data and Time Data.

Female vs Male Characters


What Percentage of Characters in Marvel Comics are Female

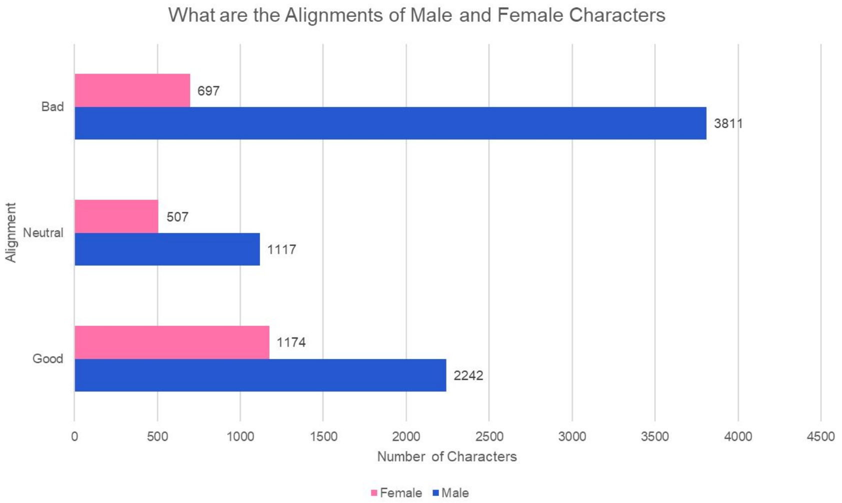

Male vs Female in the Top 100 Characters By Number of Appearances

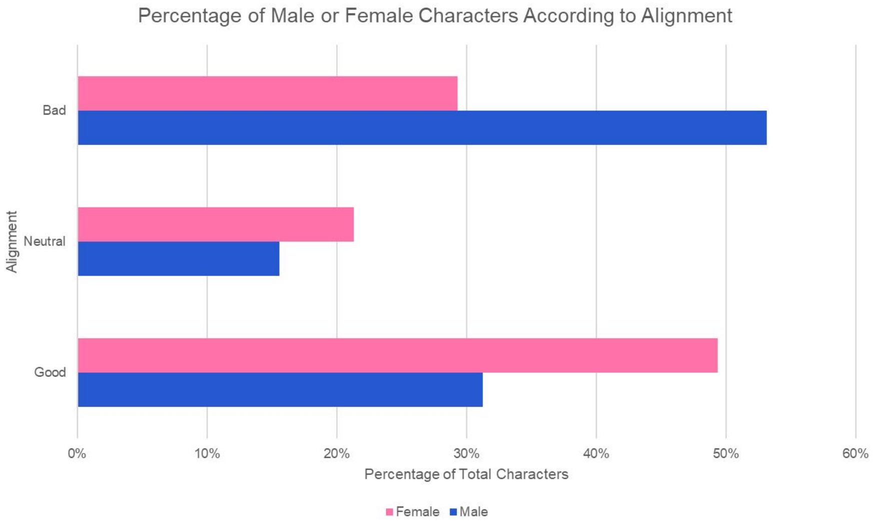
Female vs Male Characters



Female vs Male Characters

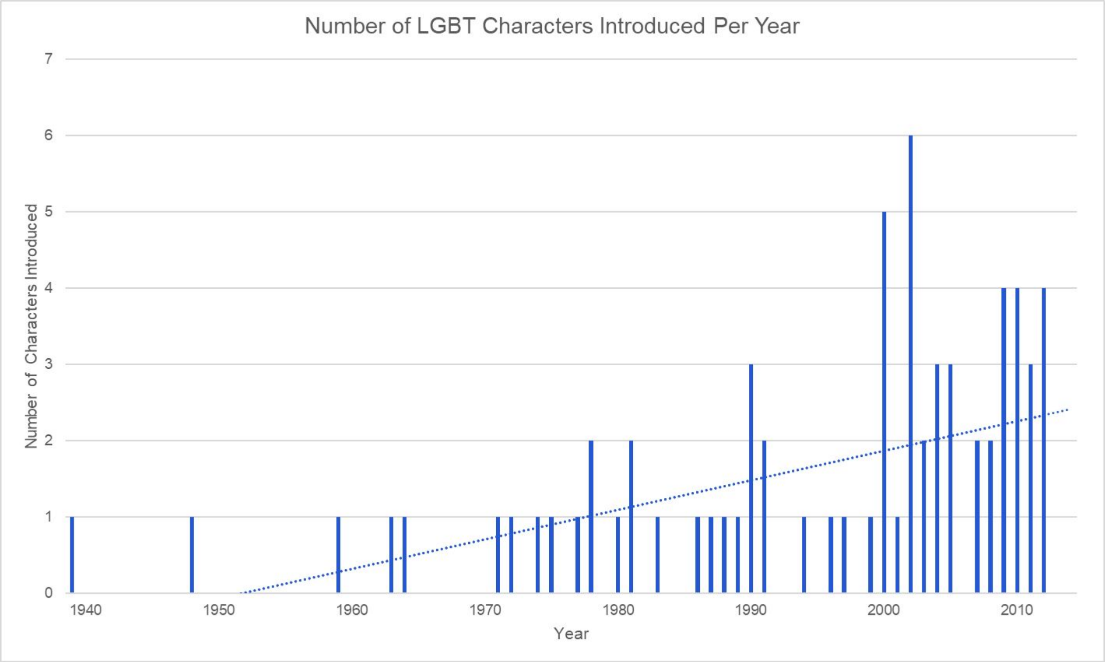
This graph makes it look like more male characters are good than female characters, however...


What are the Alignments of Male and Female Characters


Percentage of Male or Female Characters According to Alignment

When we look at the same data in terms of percentages, we can see that the *proportion* of good female characters is higher.

LGBT Characters



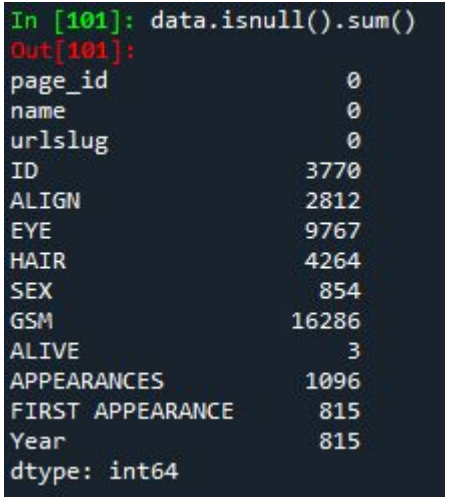
Although LGBT representation is heading in the right direction, we can see that there are still proportionally very small numbers of these characters.

Limitations and Dataset Quality

- The data only goes up as far as 2014, meaning it is almost 10 years out of date. Given that character diversity appears to be increasing in this data, I would expect that more up to date data would show a very different representation of women and LGBT characters.

- There are 6,807 rows that contained at least 1 missing value in the columns used for these questions. This means that we are losing over 41% of the data by removing these rows.

- I attempted to use classification to predict the missing values in the important columns such as identity and alignment, however I could not get it working in time.



```
In [101]: data.isnull().sum()
Out[101]:
page_id                 0
name                    0
urlslug                 0
ID                   3770
ALIGN                2812
EYE                  9767
HAIR                 4264
SEX                   854
GSM                 16286
ALIVE                   3
APPEARANCES          1096
FIRST APPEARANCE      815
Year                  815
dtype: int64
```

Conclusions

I can conclude from investigating this data that comics are increasing the number of female and LGBT characters over the years.

However there is a similar increase in male characters meaning the overall proportion of male vs female has not changed.

Future Work

In the future I would like to re-scrape the data to get a more current dataset to see if the trends seen here have continued.

I would also like to retry using classification models to predict the missing values.

If the appearances were based on time we could get a better picture of what characters are popular.