Data Analysis and Visualisation Project Report

D00230552 Conor McGuire

Introduction

Selecting a Dataset

The first step of this project was to select a dataset. I had a look through the datasets available on https://data.fivethirtyeight.com/. There were two different datasets that caught my attention. The first dataset I saw was from an article titled "Comic Books Are Still Made By Men, For Men And About Men | FiveThirtyEight". This article talks about the imbalance of genders within comic books characters, fans and writers. The second article I saw was titled "Joining The Avengers Is As Deadly As Jumping Off A Four-Story Building | FiveThirtyEight". This article looked at how many comic characters died after joining the avengers. This one was a bit too simple, so I decided to use the first dataset. While I am not a comic reader myself, I am a fan of the Marvel movies that are based on comics, so this dataset seemed interesting to me.

Metadata

The dataset was a folder containing two csv files. One named "marvel-wikia-data" and one named "dc-wikia-data". For the purposes of this project I decided to focus on just one dataset, so I chose the "marvel-wikia-data" file.

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 rows	16,376
Number of columns	13
Data Types	Includes Categorical Text Data, Numerical Data and Time Data.

Problem Statement

I want to investigate whether Marvel 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? I will investigate whether there is a difference between how men and women are represented in Marvel comics, and whether there is LGBT representation in the comics. I will look at the relationship between gender and the alignment or popularity of a character.

Data Cleaning

For the data cleaning part of the project, I decided to use Spyder, which is a python development environment. I picked this tool as we had used it in 3rd Year for investigating data so I had some experience with it already.

Importing Data

I imported the “pandas” library into my python file. This library allows me to read in the marvel-wikia-data csv file. I also imported matplotlib which allows me to create graphs with my data.

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from matplotlib.pyplot import figure


import os
os.chdir('C://Users//Conor College//OneDrive - Dundalk Institute of Technology//Documents//4th Year//4th Year Semester 2//Data Analysis and Visualisation//data-maste

data = pd.read_csv('marvel-wikia-data.csv')
```

This creates a pandas DataFrame inside Spyder. We can look at this dataframe to see the number of rows and columns.

Nam	Type	Size	Value
data	DataFrame	(16376, 13)	Column names: page_id, name, urlslug, ID, ALIGN, EYE, HAIR, SEX, GSM, ...

Next I ran some functions that allow me to get a quick overview of the data. Data.describe() gives the mean, minimum, maximum, standard deviation and more of each numerical column.

```
data.describe()
#             page_id   APPEARANCES          Year
#count   16376.000000  15280.000000  15561.000000
#mean   300232.082377     17.033377   1984.951803
#std    253460.403399     96.372959     19.663571
#min      1025.000000      1.000000   1939.000000
#25%     28309.500000      1.000000   1974.000000
#50%    282578.000000      3.000000   1990.000000
#75%    509077.000000      8.000000   2000.000000
#max    755278.000000   4043.000000   2013.000000
```

Next I ran data.isnull().sum() which shows me the total number of null rows in each column.

```
data.isnull().sum()
#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
```

This allows us to see which rows are missing the most data. This looks like gsm(gender or sexual minority) is missing a lot of rows. However when we look at the dataframe, we can see that this is because any character that is not a gender or sexual minority is given a blank cell for this column.

Female Characters	nan
Male Characters	nan
Male Characters	nan
Male Characters	nan
Male Characters	Bisexual Characters
Male Characters	nan
Male Characters	nan
Female Characters	Bisexual Characters
Male Characters	nan

We can also see that hair colour is missing 4264 rows and eye colour is missing 9767. These are the columns with the most missing data.

Cleaning Data

I can drop the urlslug and page_id columns as they are just indexes that are not useful for answering questions about the data.

```
data.drop('urlslug', axis = 1, inplace = True)
data.drop('page_id', axis = 1, inplace = True)
```

The rest of the column names are capitals. Also, one of the columns is named “align” which will cause errors as there is a function called align. I will rename all columns so that they use lowercase letters and change the name of the align column.

```
data = data.rename(columns={
    'ID':'id',
    'ALIGN':'alignment',
    'EYE':'eye',
    'HAIR':'hair',
    'SEX':'sex',
    'GSM':'gsm',
    'ALIVE':'alive',
    'APPEARANCES':'appearances',
    'FIRST APPEARANCE':'first_appearance',
    'Year':'year'})
```

As mentioned above, the eye and hair colour are missing a lot of rows. I will not need these columns in order to answer the questions I had about the data, so I can drop them from the dataset using data.drop().

```
data.drop('eye', axis = 1, inplace = True)
data.drop('hair', axis = 1, inplace = True)
```

I can also drop the first_appearance as the year column contains the same information in a better format.

```
data.drop('first_appearance', axis = 1, inplace = True)
```

This left me with 8 columns that I could use to investigate the data. I decided any rows that were missing more than half the data would not be useful, so we could drop these rows.

```
# delete rows with more than 4 missing values
missing_half = data.isnull().sum(axis=1) > 4

data.drop(data[missing_half].index, inplace = True)
```

Doing this I am left with 16,277 rows.

Running the data.isnull().sum() again, we can see how many rows are still missing data.

```
In [6]: data.isnull().sum()
Out[6]:
name              0
id             3675
alignment      2721
sex             792
gsm           16187
alive             0
appearances    1001
year            746
dtype: int64
```

Since sex, year and appearances have *relatively* low missing rows, we can drop all rows in these columns that are missing data.

```
#Drop null values in sex column
data = data.drop(data[data.sex.isnull()].index)
data['sex'].isnull().sum()  # == 0 so we know the previous line worked

data = data.drop(data[data.year.isnull()].index)
data['year'].isnull().sum()  # == 0 so we know the previous line worked

data = data.drop(data[data.appearances.isnull()].index)
data['appearances'].isnull().sum()  # == 0 so we know the previous line worked
```

Then I wanted to fix the gsm column. This column appears to be missing a lot of rows, but as mentioned above this is because any character that is not a gender or sexual minority has a blank value for this column. The gsm column lists characters that are gay, bisexual, transgender, pansexual or genderfluid. I don't need to differentiate between these values, so I will make it that there is a 1 if the character is a gsm and a 0 otherwise.

```
data['gsm'].fillna("none", inplace = True)
data['gsm']=np.where(data.gsm =="none",0,1)
```

I was then left with only the id(identity) and alignment columns having missing rows.

```
In [11]: data.isnull().sum()
Out[11]:
name             0
id            2935
alignment     2288
sex              0
gsm              0
alive            0
appearances      0
year             0
```

At this point, there are 13,954 rows remaining. This means that we have removed about 14.7% of the dataset so far. I have roughly 16,000 rows so I don't want to lose more than 16%. For this reason, I cannot drop the rows with missing values in the id and alignment column. I decided that an interesting solution would be to try and use a classification model to predict the missing values based on the other rows.

After a few attempts, I was able to create a working model that would predict these values, but I could not get it working to combine the predicted values back into the original dataset. This was very unfortunate as it meant I would lose a big chunk of data by dropping these rows. I have uploaded the full python file along with this report so you can see the attempted classification model code.

I reluctantly dropped the missing rows of the id and alignment columns so that I could begin analysing the data. I was left with 9569 rows. While I had lost a lot of data, the dataset was very large to begin with so I still had plenty to work with.

Data Exploration

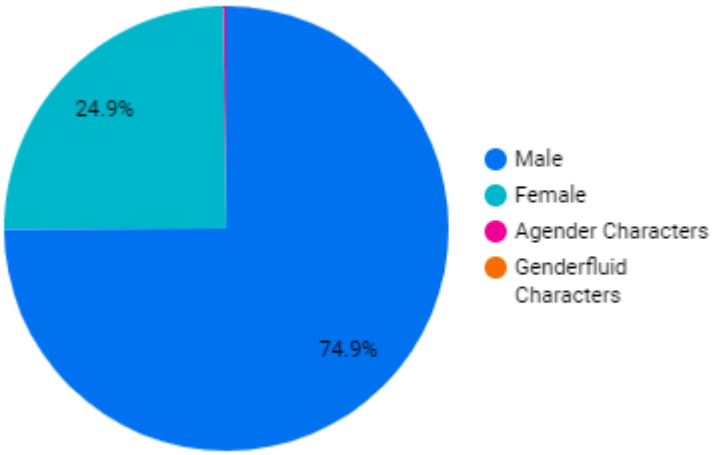
Data Visualisation Tools

To begin exploring the data, I decided to look at some simple comparisons using only one variable. The main variable I wanted to explore was the sex of characters. I originally used Spyder to write python code to generate some graphs, but I found it difficult to change the size and layout of the graphs to look nice. Now that the data was clean, I could import it into other visualisation tools. I exported the cleaned data as a csv file.

```
data.to_csv('comic_marvel_final.csv', index=False)
```

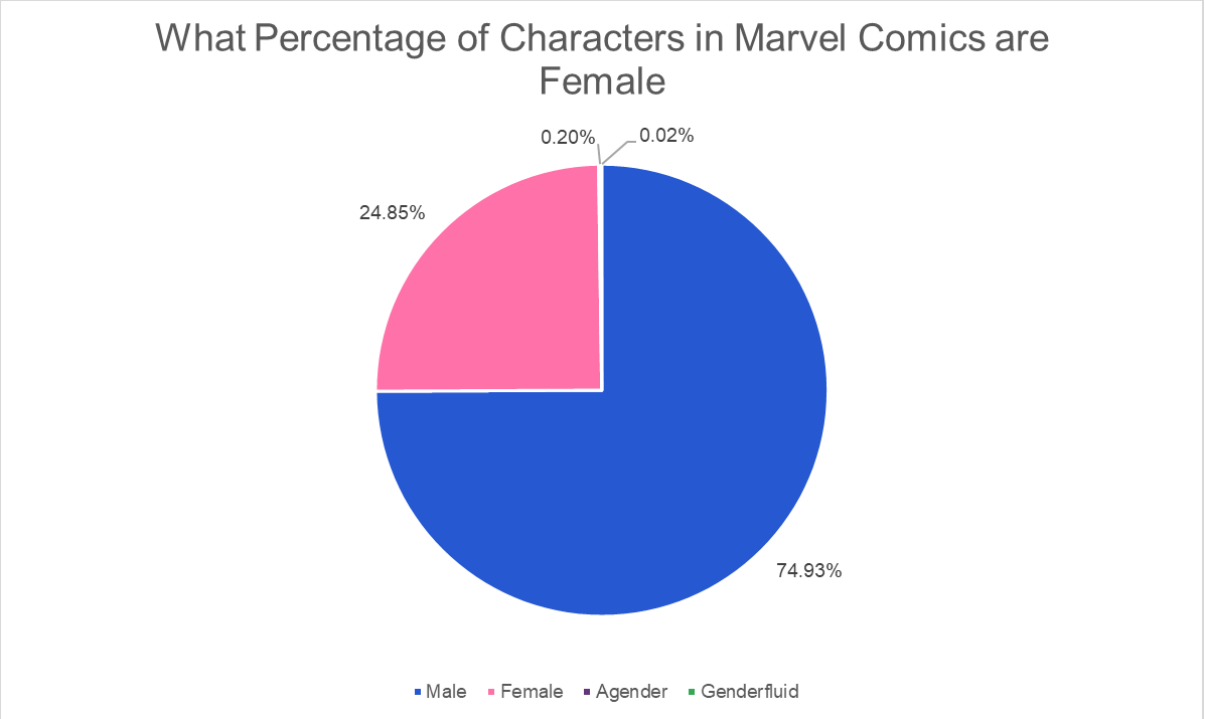
The first visualisation tool I tried was google sheets. I imported the csv and tried to create some simple charts but I didn't like how awkward it was to change the names of the values once the chart was created.

Next I tried google looker studio, this tool made it easy to import the google sheet as a data source. I was able to create a nice looking pie chart, but I was not happy as it did not show the values of the very small slices.



As you can see, there is a small fraction of agender and genderfluid characters that we cannot see percentages for. I did not like this and I could not find a way to fix it after multiple google searches.

The last tool I tried was Microsoft Excel. I imported the data in, and created the same graph again.

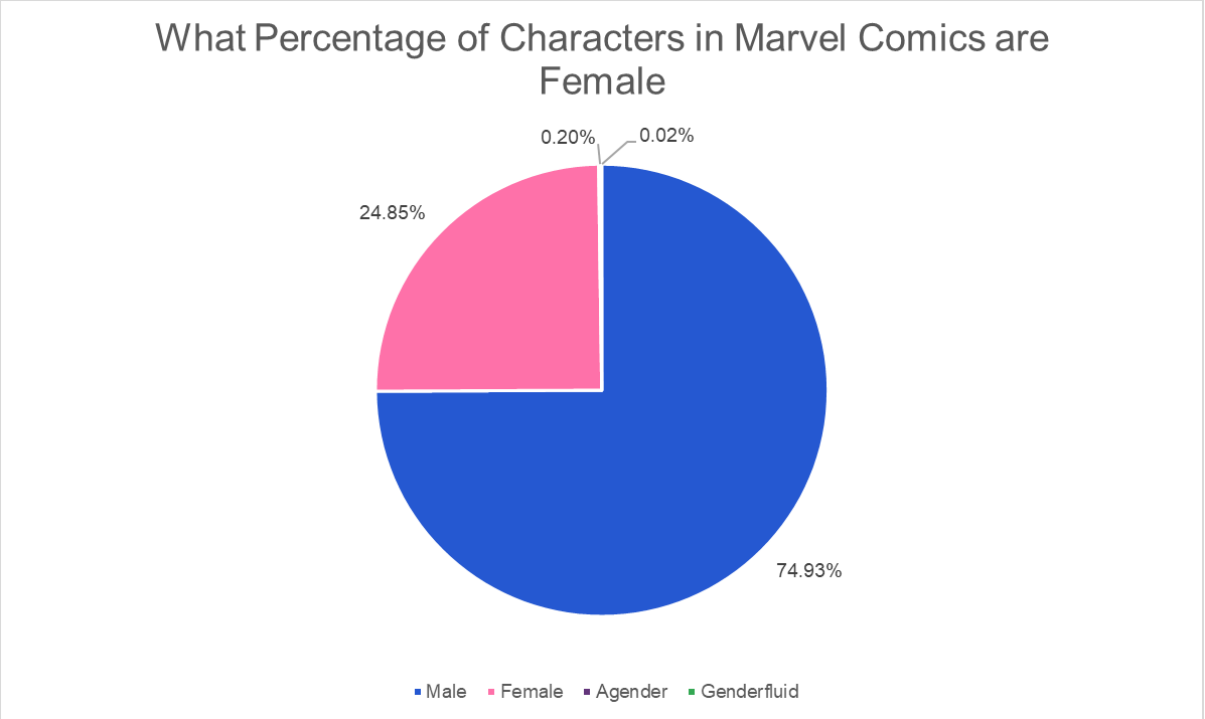


I was much happier with this tool as I could easily see the percentages of the smaller groups, as well as give titles for each chart. I decided this was the tool I would use for exploring the rest of the data.

Exploring Male vs Female

The difference between how men and women are represented in comics is the main thing I want to examine in this data. I want to find out the ratio of men to women, how this has changed over time and what are the genders of popular characters.

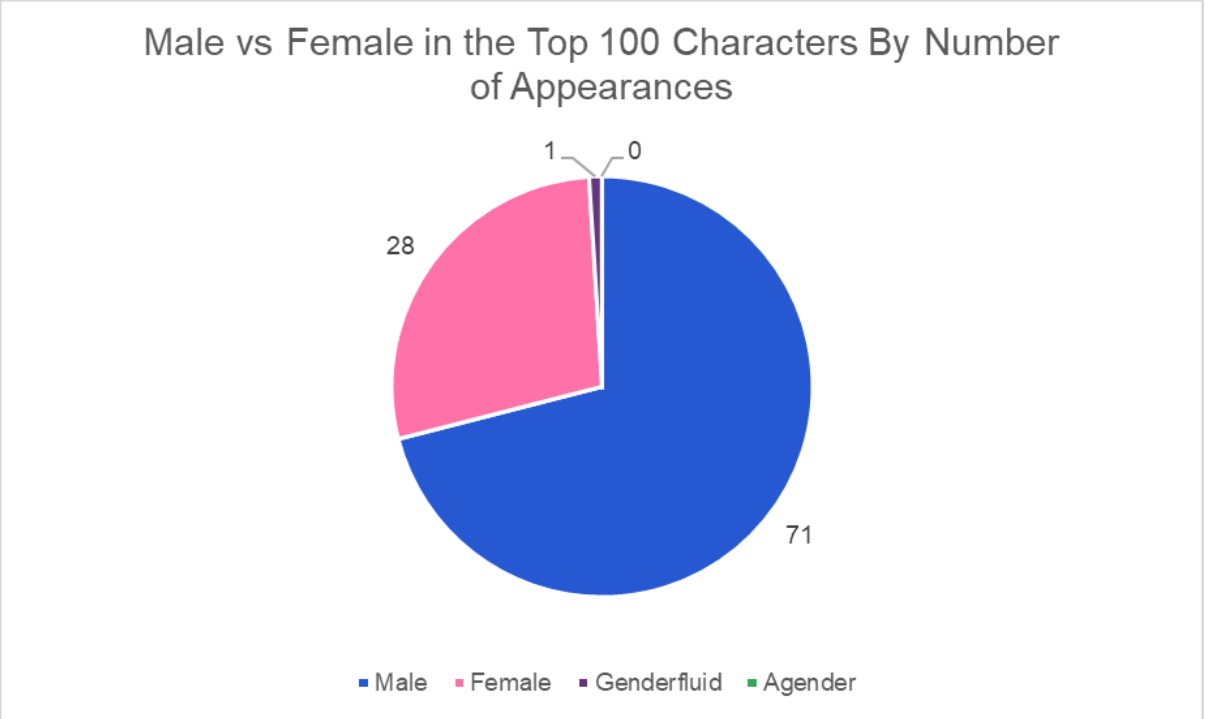
The first chart I created shows us the percentages of male, female, agender and genderfluid characters overall.



As we can see, roughly three quarters of characters are male. Roughly one quarter are female. There is a very small percentage of agender characters (0.2%) and even fewer genderfluid characters (0.02%).

This pie chart shows us that men are very over-represented in Marvel comics. There are approximately 101 men for every 100 women in the real world, according to The World Factbook. This means that the ratio is very close to 50-50, meaning that many more female characters should be introduced.

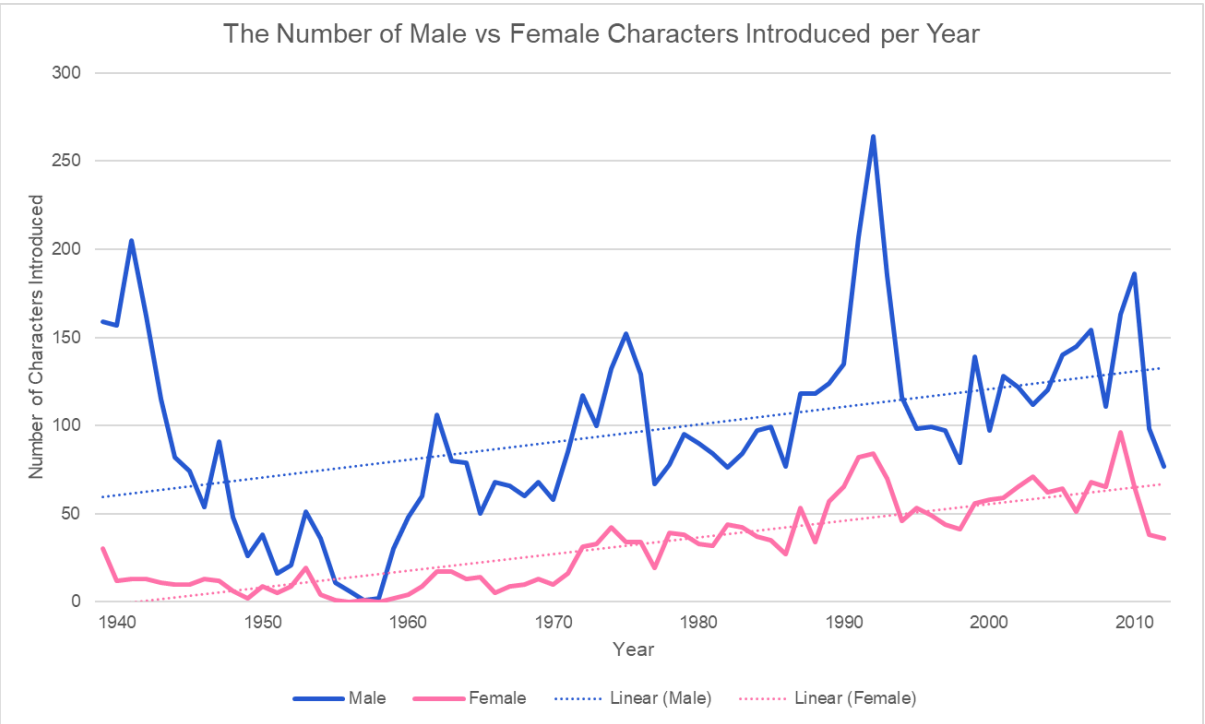
It was difficult to find a ratio of agender or genderfluid people in the real world, but I could find that approximately 1.6% of US adults are transgender or non-binary, according to About 5% of young adults in U.S. are transgender or nonbinary | Pew Research Center. This means that these demographics are also under-represented in comics.



The next thing I wanted to see was if the most popular characters are more often male or female. I looked at the top 100 characters with the highest amount of appearances. We can see from this chart that there is a very similar ratio in the top 100 as there is in the overall data. There is very slightly more female representation in the top 100 than in the overall data but not by much.

The fact that these values are similar to the first chart means that average comic readers do not lean towards one gender or another in their favourite characters.

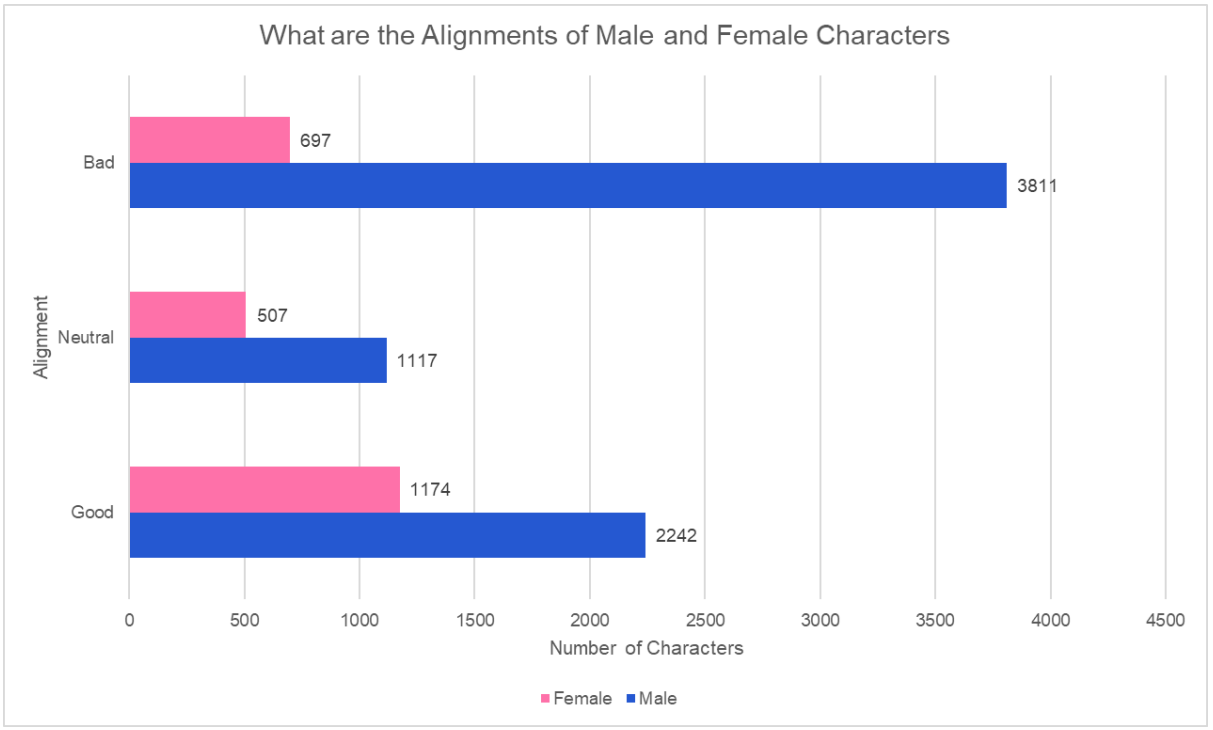
Next I investigated how many male or female characters are introduced each year.



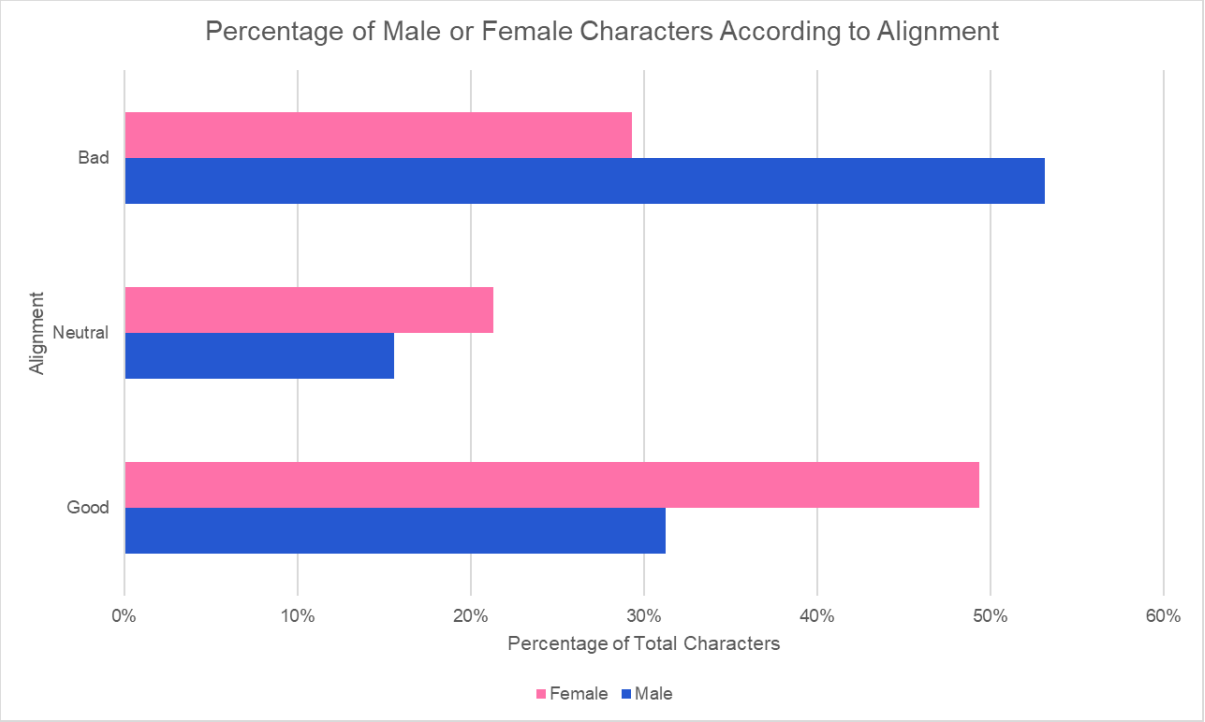
For this graph, we count the number of male characters introduced each year on the solid blue line, and female on the solid pink line. This data is still a bit unclear, so I added a **trendline** for each gender. This shows us that the comics are adding more female and male characters at roughly the same rate.

Because both male and female characters are increasing over time, the 75-25 split from the pie chart is unlikely to change. If the comic writers wanted to balance the number of men and women in comics, they would need to add more women than men, so that the values are eventually equal.

The final male-female comparison I wanted to see was the split between alignment. This dataset shows alignment as good, neutral or bad. This tells us whether a character is a “hero” or a “villain” in the comics.



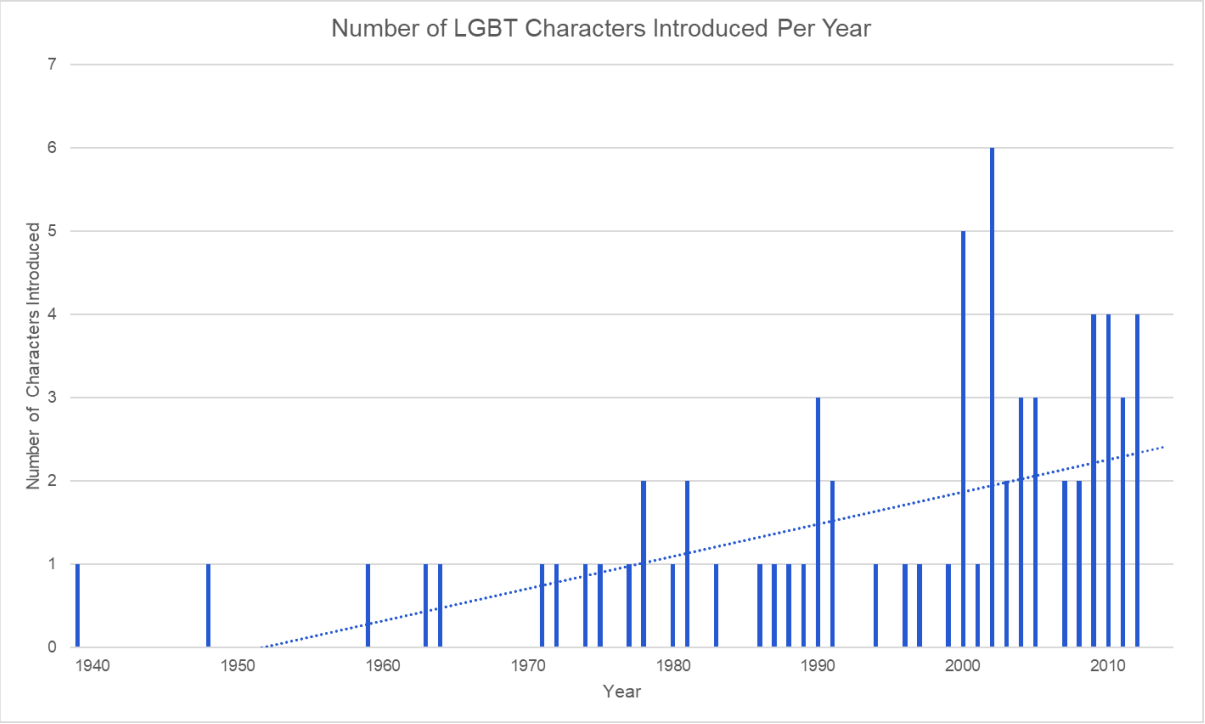
The first chart shows the number of men and women that are good, neutral and bad. I didn’t like this way of visualising the information. Since there are less female characters overall, it will look like there are less good female characters than male. However when we look at percentages instead, we can get a different impression of the data.



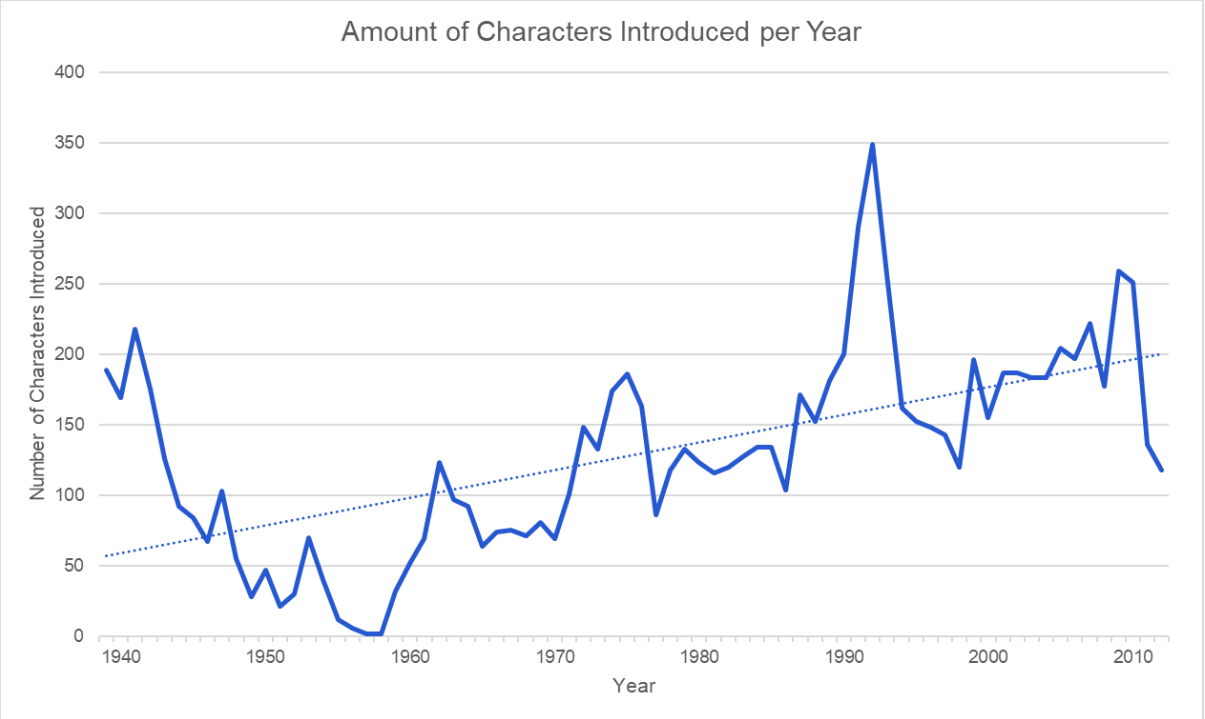
From this chart we can see that even though there are less women characters, a higher *percentage* of those characters are good, almost 50%. We can see that there are almost double the percentage of bad male characters than bad female characters.

This shows us that women are treated very differently than men when it comes to alignment.

Exploring LGBT Representation



Looking at how many LGBT characters are introduced each year, we can see that the number is increasing. However if we look at the y-axis we can see that the maximum number of LGBT characters introduced in a year was 6 in 2002. This is still far lower than the number of non-LGBT characters. The total number of LGBT characters is 68.



When we look at the total number of characters introduced per year, we can see that in 2002 there were roughly 200 characters introduced. 6 LGBT characters out of a total 200 would mean 3% of characters introduced that year were LGBT. This article: LGBT Identification in U.S. Ticks Up to 7.1% tells us that LGBT identification was 7% in 2022. This means the comic writers should introduce more LGBT characters if they want to represent all groups fairly.

Data Limitations and Quality

There were some major limitations with this data. Firstly, there were very many missing rows which meant a lot of the data was lost in the cleaning process. I went from 16,376 rows to 9,569. This is a loss of 41.5%. This would normally make the data unusable but since we had so many rows to begin with, we are still ok but it is not an ideal situation. If I had gotten the classification model to work correctly I would only lose 14.7% through the cleaning process.

The other main issue with this data is that it is very old. It was scraped from the Marvel Wikia in August 2014. The Marvel movies have exploded in popularity since then and I expect this would have driven a lot of change in modern comics. The issues of gender and sexuality are even more talked about today than they would have been when this data was originally scraped. This means that there could be a big difference in the data from 2014 to now.

One small issue that could be considered is the “appearances” column. For one of my charts, I used the number of appearances to select the top 100 most popular characters.

However since Marvel comics have been going since the 1940s, some characters are far older than others. This could mean that an old character could have more appearances than an newer character even if the newer one is more popular with readers, just because of how much time there has been since the older character has been introduced. This means that appearances is not a perfect way to measure “popularity”.

Looking into diversity, it would be also interesting to see if we updated the dataset to include variables such as race or religions of characters to see if there is fair representation in the comics. There could be some issues with this however, as many of the characters in the comics are not human so race and religion will not apply to them as easily as gender or alignment.

Conclusion

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. If the writers of the comics want to increase the diversity of the characters, they should increase how many new female characters they add in order to balance out the genders. They should also add some more LGBT characters per year to increase diversity there also. As I mentioned before, the Marvel movies are becoming extremely popular in recent years which is encouraging many people to start reading the comics. I think that if the comic writers increase the diversity of their characters, the comics will appeal to a wider audience of potential new readers.

As I mentioned earlier, the data could be further explored if it was re-scraped from the wiki to include more up-to-date information and perhaps new variables such as race and religion. This data could help Marvel comic writers decide what are the new kinds of characters they should introduce.