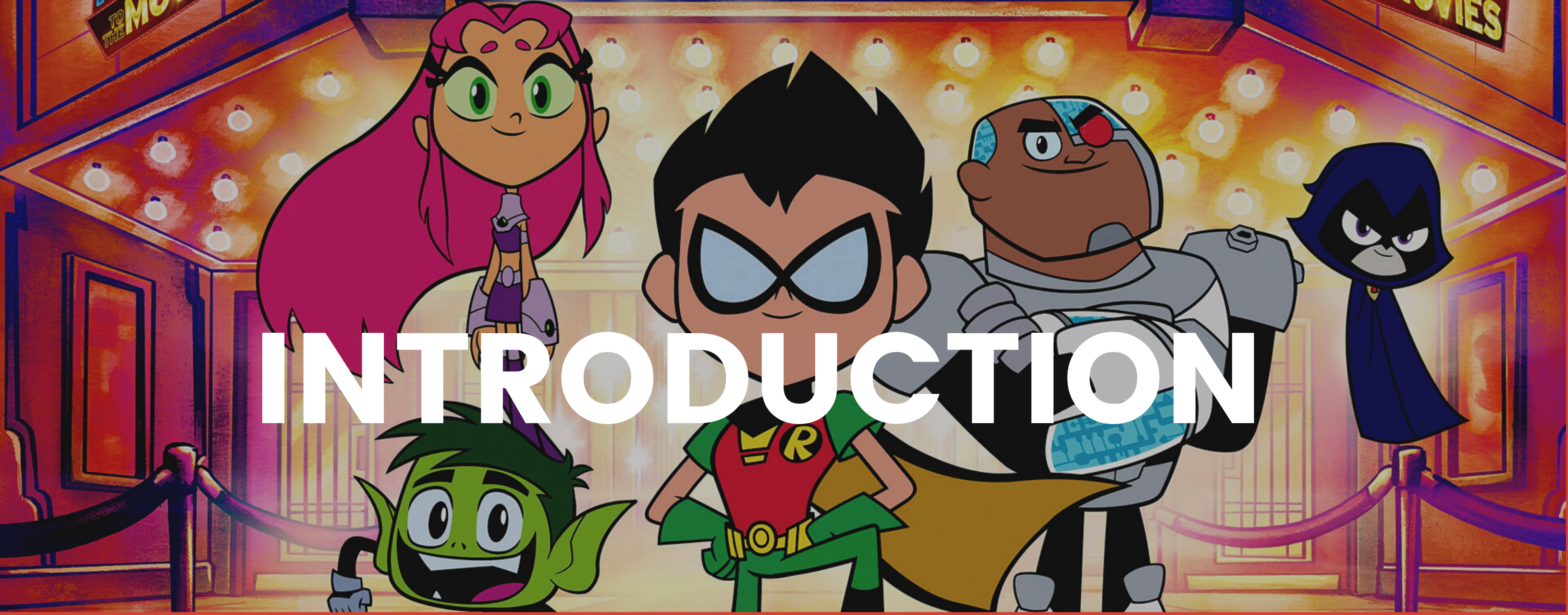
DIVERSITY IN DC COMICS

Kelly Dang

EXECUTIVE SUMMARY

There are few things more iconic than DC comics. DC has transcended itself as an essential part of everyones' childhood. A symbol of freedom, strength, and resilience. This report dives into DC comic books, examining a data set with information on DC characters alongside multiple attributes of those characters. The variables of interest are appearances, sex, eye color, hair color, identity (secret vs. public vs. unknown), and alignment (good vs. bad vs neutral). The main question was if there was a relationship between these categorical variables and appearances and if there could be a linear model curated to help predict the number of appearances. A linear regression was created with the categorical variables as the x variables and the number of appearances as the y variables. The linear regression model produced coefficients for each attribute. With t-values and p-values in mind, only 3 attributes were statistically significant including good characters, male characters and reformed criminals. Additionally, descriptive statistics took place that showed the distribution of each categorical variable as well as other graphs incorporating time and proportions. As it turns out there are still not as many female characters as male characters in the DC comic book work. Additionally, most characters have blue eyes, a feature predominantly described as eurocentric. It is important to note that this model did not take into account that a comic book character may appear more frequently because he/she was introduced earlier. This would be grounds for later research and analysis. Upon further secondary research, it is noted that diversity is still something that the DC franchise needs to work on, but it is slowly but surely getting there through becoming more inclusive and representative of today's current landscape. With all that being said, there are three recommendations from this project - DC needs to have more female characters, DC needs to have more characters from different ethnic backgrounds, and finally DC needs to have more characters that represent different aspects of the LGBTQ community. Everyone deserves a superhero that he or she can relate to and hopefully one day we can all get to that point.



INTRODUCTION

Growing up, it seemed like no one could escape the power of DC. Thank you Malcolm Wheeler-Nicholson for that. No matter how old you were or how in tune or out of tune with pop culture you were, you knew who Superman and Wonderwoman were. DC was and still is ubiquitous across the globe and for good reason. We all want someone to root for, someone that reminds us that at the end of the day, good will always triumph over evil and that hope prevails no matter the circumstance.

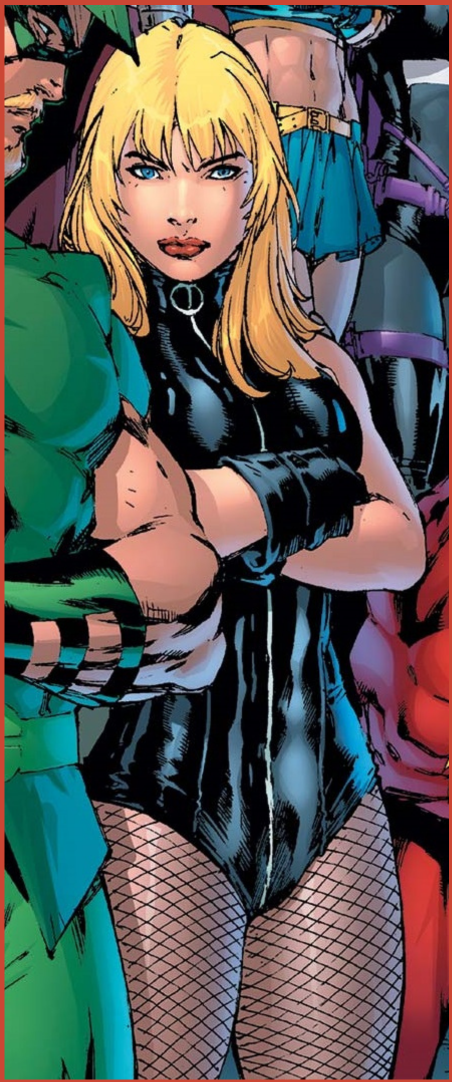
There have been so many moments of my childhood connected with DC. I couldn't walk down the toy aisle as a kid without checking out the superhero action figures. I also couldn't resist the urge to watch Teen Titans Go on Cartoon Network on lazy Saturdays instead of doing my homework.. This adoration spread to the rest of my family. My younger brother, Brandon became the biggest DC fan and his closet filled to the brim with DC shirts. I bet if I called him right now, he would be playing the video game Injustice.

However, DC had more of a profound impact than I could've ever imagined as a kid. When my aunt from Vietnam first came to the USA, the first movie she ever saw in theatres was Superman Returns. Because of the film, Superman became synonymous with America to her. He represented strength, resilience, hope, and the American dream, something she hoped her kids would be able to achieve when they got older. Because DC has been connected with my life, I wanted to look more into the comic books. We all have watched superhero movies, but not many of us have dived into the world of reading the comics. The movies are just the tip of the iceberg, as comic books offer thousands of characters, each with a unique story to offer readers.

THE DATA

I was able to find this cool dataset on Kaggle from FiveThirtyEight. The data came from the webpage, DC Wikia, a website dedicated to all things DC related. The information was scraped in 2014, with characters being scraped on August 24, appearance counts on September 2, and the month and year of the first issue each character appeared in was scraped on October 6. For this dataset, there are 13columns labeled page_id, name, urlslug, ID, align, eye, sex, GSM, alive, appearance, first appearence, and year. Inserted below is a table with the variable names and their definitions.

Variable	Definition
Page_id	The unique identifier for that characters page within the wikia page
Name	The name of the character
urlslug	The unique url within the wikia that takes you to the character
ID	The identity status of the character (Secret Identity, Public Identity)
Align	If the character is Good, Bad, or Neural
Eye	Eye color of the character
Hair	Hair color of the character
Sex	Sex of the character (male, female, genderless)
GSM	If the character is a gender or sexual minority (e.g. homosexual characters, bisexual characters)
Alive	If the character is alive or deceased
Appearances	The number of appearances of the character in comic books (as of Sep 2. 2014. Number will become increasingly out of date as time goes on.)
First Appearance	The month and year of the character's first appearance in a comic book, if available
Year	The year of the character's first appearance in a comic book, if available



The column that most intrigued me was the appearances column, some characters were featured considerably more than others. That got me thinking if I could create a model that could compute the approximate number of appearances of a character based on the categorical variables of the identity status (secret vs. public), alignment(good, bad, neutral), eye color, hair color, and sex of the DC comic book charcters. I planned to use a linear regression to examine this relationship because the dependent variable was continuous and not categorical.

I had a multitude of questions I wanted to examine. I wondered about the proportion of male characters vs female characters as the comic book world has always catered more toward men and is male centric. I wanted to see if there was a shift in female characters as time has gone by. I also wondered if eurocentric features would be featured more prominently such as blue eyes and blond hair.

I ended up filtering out missing data in my excel sheet, this led to a significant decrease in the data. I had 2180 rows of data as opposed to the original 6897.

```{r}
data <- read.csv("~/DC File.csv")
```

```{r}
head(data,10)
```

Description: df [10 x 13]

	ï..page_id <int>	name <chr>	urlslug <chr>	ID <chr>	ALIGN <chr>	EYE <chr>	HAIR <chr>
1	1422	Batman (Bruce Wayne)	\\/wiki\\/Batman_(Bruce_Wayne)	Secret Identity	Good Characters	Blue Eyes	Black Hair
2	23387	Superman (Clark Kent)	\\/wiki\\/Superman_(Clark_Kent)	Secret Identity	Good Characters	Blue Eyes	Black Hair
3	1458	Green Lantern (Hal Jordan)	\\/wiki\\/Green_Lantern_(Hal_Jordan)	Secret Identity	Good Characters	Brown Eyes	Brown Hair
4	1659	James Gordon (New Earth)	\\/wiki\\/James_Gordon_(New_Earth)	Public Identity	Good Characters	Brown Eyes	White Hair
5	1576	Richard Grayson (New Earth)	\\/wiki\\/Richard_Grayson_(New_Earth)	Secret Identity	Good Characters	Blue Eyes	Black Hair
6	1448	Wonder Woman (Diana Prince)	\\/wiki\\/Wonder_Woman_(Diana_Prince)	Public Identity	Good Characters	Blue Eyes	Black Hair
7	1486	Aquaman (Arthur Curry)	\\/wiki\\/Aquaman_(Arthur_Curry)	Public Identity	Good Characters	Blue Eyes	Blond Hair
8	1451	Timothy Drake (New Earth)	\\/wiki\\/Timothy_Drake_(New_Earth)	Secret Identity	Good Characters	Blue Eyes	Black Hair
9	71760	Dinah Laurel Lance (New Earth)	\\/wiki\\/Dinah_Laurel_Lance_(New_Earth)	Public Identity	Good Characters	Blue Eyes	Blond Hair
10	1380	Flash (Barry Allen)	\\/wiki\\/Flash_(Barry_Allen)	Secret Identity	Good Characters	Blue Eyes	Blond Hair

1-10 of 10 rows | 1-8 of 13 columns

Description: df [10 x 13]

ID <chr>	ALIGN <chr>	EYE <chr>	HAIR <chr>	SEX <chr>	GSM <chr>	ALIVE <chr>	APPEARANCES <int>	FIRST.APPEARANCE <chr>	YEAR <int>
Secret Identity	Good Characters	Blue Eyes	Black Hair	Male Characters		Living Characters	3093	1939, May	1939
Secret Identity	Good Characters	Blue Eyes	Black Hair	Male Characters		Living Characters	2496	1986, October	1986
Secret Identity	Good Characters	Brown Eyes	Brown Hair	Male Characters		Living Characters	1565	1959, October	1959
Public Identity	Good Characters	Brown Eyes	White Hair	Male Characters		Living Characters	1316	1987, February	1987
Secret Identity	Good Characters	Blue Eyes	Black Hair	Male Characters		Living Characters	1237	1940, April	1940
Public Identity	Good Characters	Blue Eyes	Black Hair	Female Characters		Living Characters	1231	1941, December	1941
Public Identity	Good Characters	Blue Eyes	Blond Hair	Male Characters		Living Characters	1121	1941, November	1941
Secret Identity	Good Characters	Blue Eyes	Black Hair	Male Characters		Living Characters	1095	1989, August	1989
Public Identity	Good Characters	Blue Eyes	Blond Hair	Female Characters		Living Characters	1075	1969, November	1969
Secret Identity	Good Characters	Blue Eyes	Blond Hair	Male Characters		Living Characters	1028	1956, October	1956

1-10 of 10 rows | 5-14 of 13 columns

DESCRIPTIVE STATISTICS

I wanted to get distributions of all the categorical variables. I decided to create these particular graphs in R and Tableau. For R, I made sure I loaded in all of the appropriate packages I needed.

````
```{r}
install.packages("tidyverse")
library(tidyverse)
```
````

````
```{r}
library(ggplot2)
library(ggthemes)
```
````

````
```{r}
library(dplyr)
```
````

````
# Reorder Function - this was to make sure the graphs I would create in R would be in descending order
```{r}
reorder_size <- function(x) {
    factor(x, levels =names(sort(table(x), decreasing = TRUE)))
}
```
````

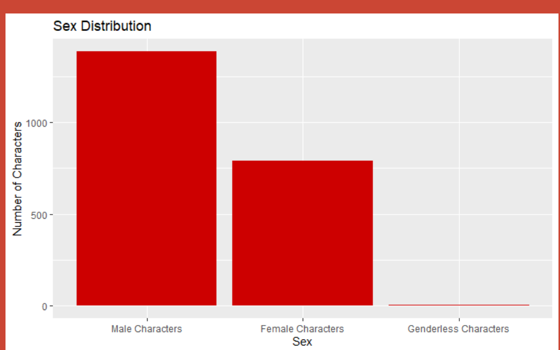
MALE VS. FEMALE

````
```{r}
table(data$SEX)
table1 <- table(data$SEX)
prop.table(table1)
```
````

Female Characters	Genderless Characters	Male Characters
789	3	1387

Female Characters	Genderless Characters	Male Characters
0.362092703	0.001376778	0.636530519

````
```{r}
data %>%
    ggplot(aes(x = reorder_size (SEX))) +
          geom_bar(fill="red3") +
          labs(title = "Sex Distribution",
               x = "Sex",
               y = "Number of Characters",)
```
````



To no surprise, only 36% of characters are female while 63% are male. Also to no surprise, a majority of comic book characters have blue eyes even though only 8% of the population has them and they are predominately referred to as a eurocentric trait. Superman, Batman, and Wonder Woman, the 3 biggest stars of the DC universe, all have blue eyes and were introduced early on in 1938, 1939, and 1941.

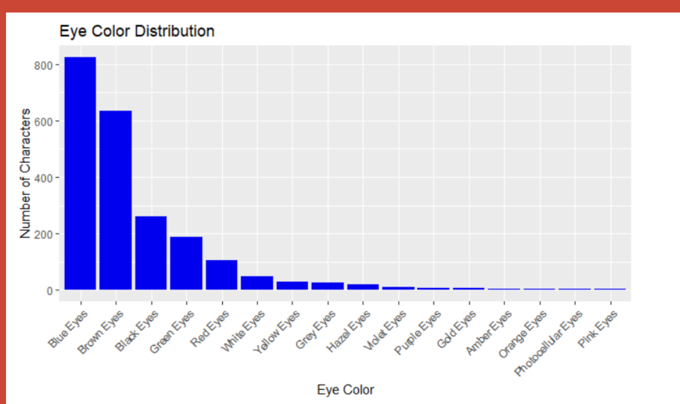
EYE COLOR

````
```{r}
table(data$EYE)
```
````

Amber Eyes	Black Eyes	Blue Eyes	Brown Eyes
3	262	824	633
Orange Eyes	**Photocellular Eyes**	**Pink Eyes**	**Purple Eyes**
3	3	3	8

Gold Eyes	Green Eyes	Grey Eyes	Hazel Eyes
6	189	28	20
Red Eyes	**Violet Eyes**	**White Eyes**	**Yellow Eyes**
107	10	50	30

````
```{r}
data %>%
    ggplot(aes(x = reorder_size (EYE))) +
         geom_bar(fill = "blue2") +
         labs(title = "Eye Color Distribution",
              x = "Eye Color",
              y = "Number of Characters",) +
         theme(axis.text.x = element_text(angle = 45, hjust=1))
```
````



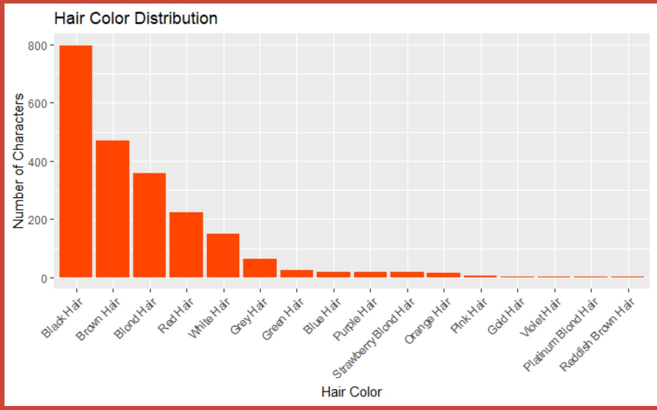
HAIR COLOR

````
```{r}
table(data$HAIR)
```
````

```
  Black Hair          Blond Hair             Blue Hair            Brown Hair
         798                 359                    19                   470
 Orange Hair           Pink Hair   Platinum Blond Hair           Purple Hair
          15                   5                     2                    18
Violet Hair          White Hair
           4                 151
```

```
 Gold Hair               Green Hair           Grey Hair
         4                       26                  64
  Red Hair   Reddish Brown Hair Strawberry Blond Hair
       224                    2                    18
```

````
```{r}
data %>%
   ggplot(aes(x = reorder_size (HAIR))) +
        geom_bar(fill="orangered1") +
        labs(title = "Hair Color Distribution",
             x = "Hair Color",
             y = "Number of Characters",) +
        theme(axis.text.x = element_text(angle = 45, hjust=1))
```
````



The distribution here better mirrors the real world, with a majority having black or brown hair, then blonde, and then red. Here, eurocentric features are not favored.

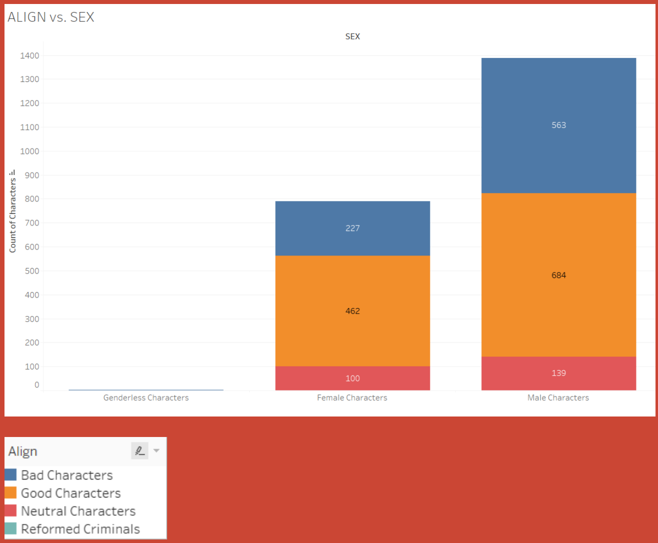
ALIGNMENT

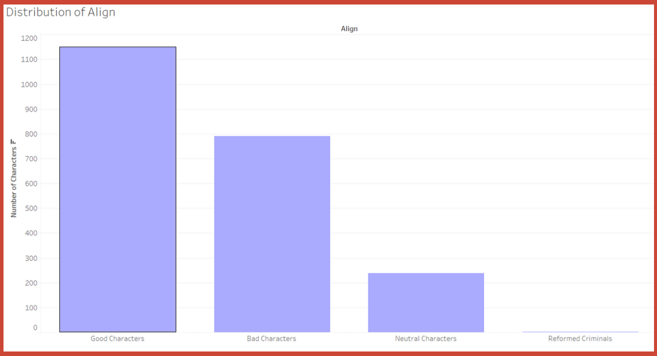
Note from here on, R was loaded into Colab due to technical issus with the software on my laptop. Tableau was used here be cause the R bar chart did not show a value for reformed criminals because of how small the value was.

```
%%R
table(data$ALIGN)
```

```
    Bad Characters    Good Characters Neutral Characters Reformed Criminals
               791               1148                239                  1
```

```
%%R
data %>%
  ggplot(aes(x = reorder_size (ALIGN))) +
      geom_bar(fill="purple") +
      labs(title = "Alignment Distribution",
           x = "Alignment",
           y = "Number of Characters",)
```





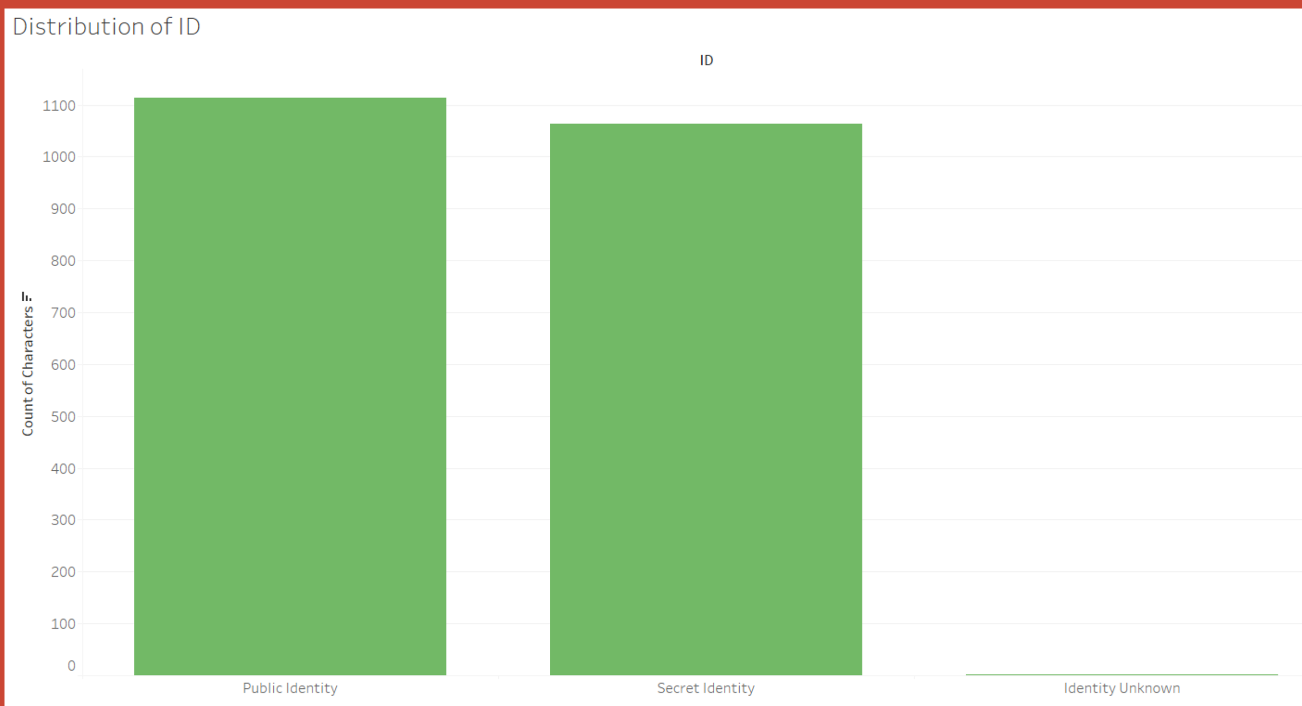
The distribution illustrates that a majority of characters are good characters with only 1 reformed character. I found that there is a bigger difference proportionally between good to bad characters for women than for men. Both genders however have more good characters than bad. The one genderless character is a bad character.

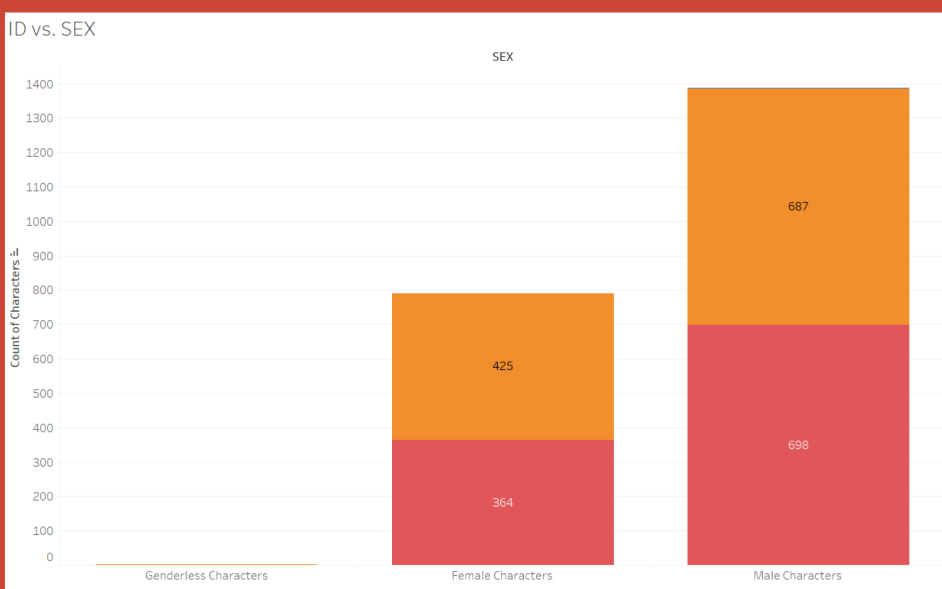
IDENTITY

```
%%R
table(data$ID)
```

```
Identity Unknown  Public Identity  Secret Identity
               2             1113             1064
```

```
%%R
data %>%
  ggplot(aes(x = reorder_size (ID))) +
      geom_bar(fill="green") +
      labs(title = "ID Distribution",
          x = "ID",
          y = "Number of Characters",)
```





These distributions illustrate that a most characters have a public identity, but not by much. The genderless character has a public identity. I wanted to dive into alignment relative to gender and found that there is a bigger difference proportionally between public to secret identity for women than for men. It is almost evenly split between the two for men. Both genders however have more public identities than secret proportionally.

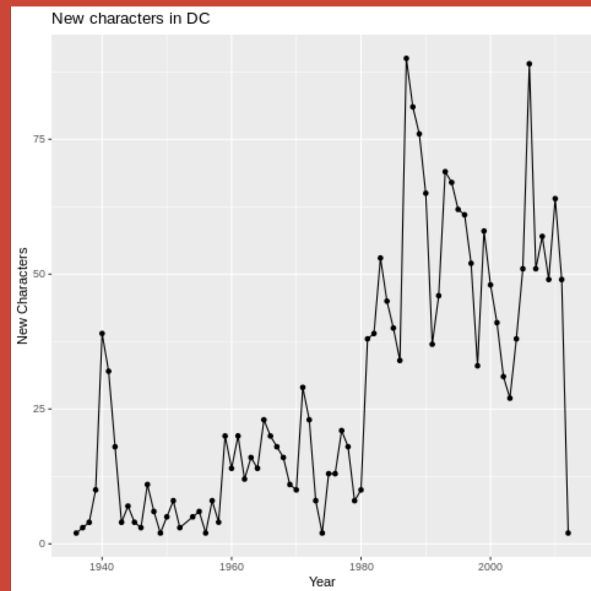
ID

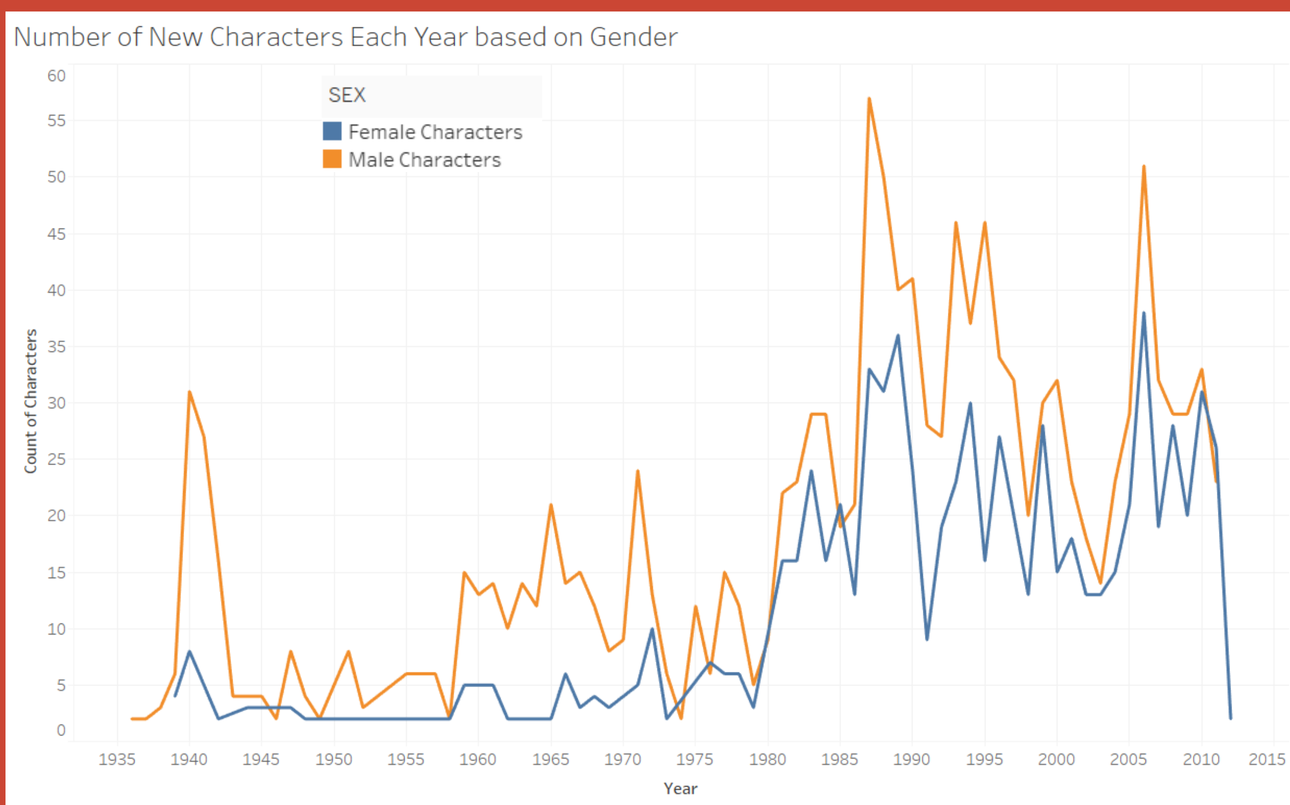
- Identity Unknown
- Public Identity
- Secret Identity

APPEARANCES THROUGHOUT THE YEARS

```
%%R
yearly_new_characters <- data %>%
    group_by(YEAR) %>%
    summarize(num = n())
```

```
%%R
yearly_new_characters %>%
    ggplot(aes(x = YEAR, y = num)) +
        geom_line() +
        geom_point() +
        labs(title = "New characters in DC",
             x = "Year",
             y = "Number of New Characters",)
```





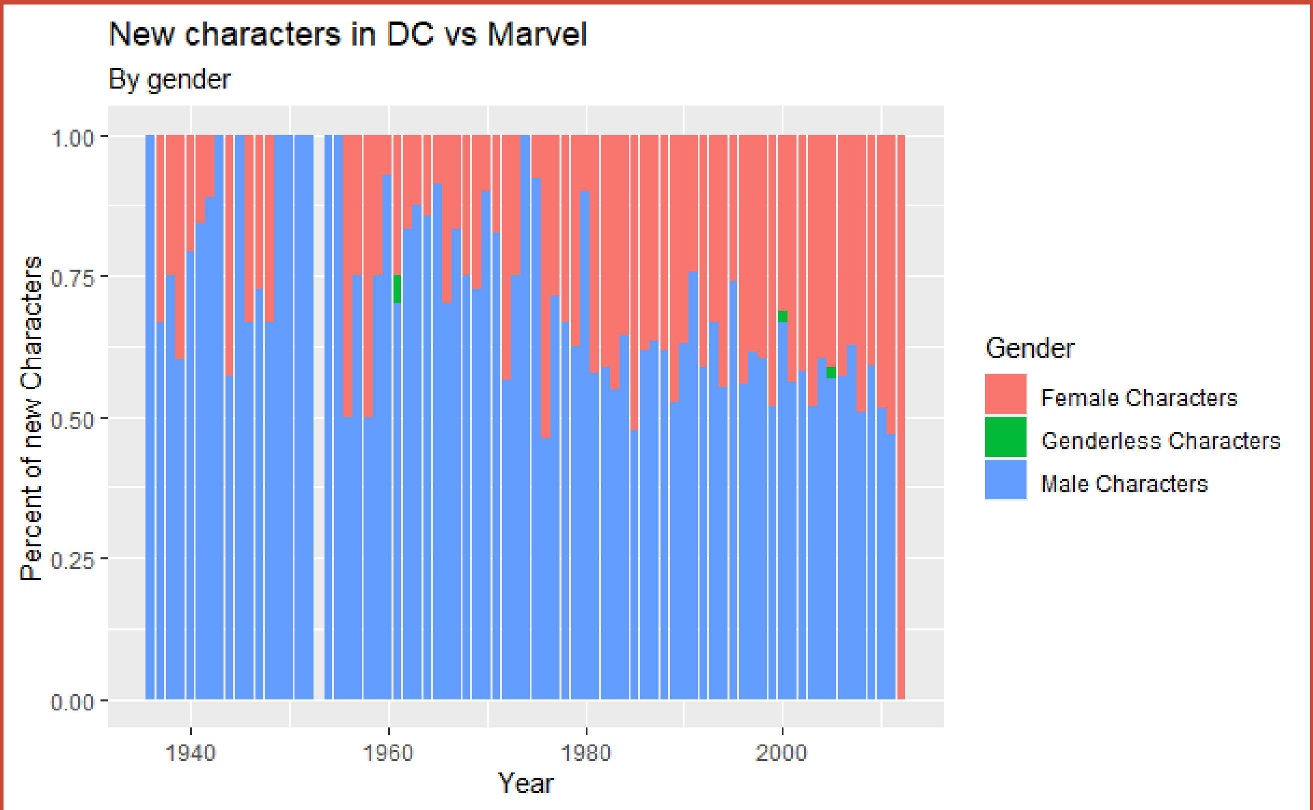
Next, I wanted to look at the number of new characters introduced every year. There seems to be a lot of ups and downs with the data and no clear pattern. Because of this I decided to separate the characters into male and females. Most of the time there are more male characters introduced than female every year. There are a couple exceptions in the data, for example in the years 1976 and 1985 there were more female characters introduced.

APPEARANCES THROUGHOUT THE YEARS CONT'D

```
%%R
characters_by_sex <- data %>%
    filter(!is.na(SEX)) %>%
    group_by(YEAR, SEX) %>%
    summarize(num = n())
```

```
`summarise()` has grouped output by 'YEAR'. You can override using the `.groups` argument.
```

```
%%R
characters_by_sex %>%
    ggplot(aes(x = YEAR, y = num, fill = SEX)) +
        geom_bar(stat = "identity", position = "fill") +
        labs(title = "New characters in DC vs Marvel",
             subtitle = "By gender",
             x = "Year",
             y = "Percent of new Characters",
             fill = "Gender")
```



Continuing on from the last graph, I wanted to illustrate the same information as before, but in a different manner. For this graph, I compared new female vs. male characters introduced through percentages instead. I did this because it is easier to compare the two genders on a relative basis.

LINEAR REGRESSION

In order to run a linear regression to see the relationship between the categorical variables and number of appearances, I used this code which helped to dummy code, output the linear regression with coefficients, and give a summary of the data.

INPUT

```
```{r}
levels_ID <- levels(factor(data$ID))
print(levels_ID)
levels_ALIGN <- levels(factor(data$ALIGN))
print(levels_ALIGN)
levels_EYE <- levels(factor(data$EYE))
print(levels_EYE)
levels_HAIR <- levels(factor(data$HAIR))
print(levels_HAIR)
levels_SEX <- levels(factor(data$SEX))
print(levels_SEX)
regression <- lm(data$APPEARANCES ~ data$ID + data$ALIGN + data$EYE + data$HAIR
                + data$SEX)
summary(regression)
```
```

OUTPUT

```
[1] "Identity Unknown" "Public Identity"  "Secret Identity"
[1] "Bad Characters"     "Good Characters"     "Neutral Characters" "Reformed Criminals"
 [1] "Amber Eyes"         "Black Eyes"          "Blue Eyes"
 [4] "Brown Eyes"         "Gold Eyes"           "Green Eyes"
 [7] "Grey Eyes"          "Hazel Eyes"          "Orange Eyes"
[10] "Photocellular Eyes" "Pink Eyes"           "Purple Eyes"
[13] "Red Eyes"           "Violet Eyes"         "White Eyes"
[16] "Yellow Eyes"
 [1] "Black Hair"            "Blond Hair"            "Blue Hair"
 [4] "Brown Hair"            "Gold Hair"             "Green Hair"
 [7] "Grey Hair"             "Orange Hair"           "Pink Hair"
[10] "Platinum Blond Hair"   "Purple Hair"           "Red Hair"
[13] "Reddish Brown Hair"    "Strawberry Blond Hair" "Violet Hair"
[16] "White Hair"
[1] "Female Characters"     "Genderless Characters" "Male Characters"

Call:
lm(formula = data$APPEARANCES ~ data$ID + data$ALIGN + data$EYE +
    data$HAIR + data$SEX)

Residuals:
    Min      1Q  Median      3Q     Max
-131.86  -46.90  -20.47    8.13 2977.41
```

```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                    23.377    127.380   0.184   0.8544    
data$IDPublic Identity         17.844     98.841   0.181   0.8568    
data$IDSecret Identity         38.205     98.821   0.387   0.6991    
data$ALIGNGood Characters      51.521      6.842   7.530 7.51e-14 ***
data$ALIGNNeutral Characters   23.288     10.694   2.178   0.0295 *  
data$ALIGNReformed Criminals   28.582    139.532   0.205   0.8377    
data$EYEBlack Eyes            -63.913     81.053  -0.789   0.4305    
data$EYEBlue Eyes             -14.485     80.773  -0.179   0.8577    
data$EYEBrown Eyes            -56.107     80.712  -0.695   0.4870    
data$EYEGold Eyes             -67.277    101.747  -0.661   0.5085    
data$EYEGreen Eyes            -25.535     81.333  -0.314   0.7536    
data$EYEGrey Eyes             -33.436     84.832  -0.394   0.6935    
data$EYEHazel Eyes            -20.541     86.423  -0.238   0.8122    
data$EYEOrange Eyes           -73.467    128.646  -0.571   0.5680    
data$EYEPhotocellular Eyes    -18.236    113.798  -0.160   0.8727    
data$EYEPink Eyes             -38.005    114.284  -0.333   0.7395    
data$EYEPurple Eyes            66.183     95.273   0.695   0.4873    
data$EYERed Eyes              -43.084     81.971  -0.526   0.5992    
data$EYEViolet Eyes           -69.167     96.681  -0.715   0.4744    
data$EYEWhite Eyes            -64.299     83.061  -0.774   0.4389    
data$EYEYellow Eyes           -51.160     84.910  -0.603   0.5469    
data$HAIRBlond Hair           -15.338      9.775  -1.569   0.1168    
data$HAIRBlue Hair            -22.215     35.983  -0.617   0.5371    
data$HAIRBrown Hair            -9.075      8.353  -1.086   0.2774    
data$HAIRGold Hair            -21.410     76.759  -0.279   0.7803    
data$HAIRGreen Hair            34.241     28.655   1.195   0.2322    
data$HAIRGrey Hair            -24.991     18.812  -1.328   0.1842    
data$HAIROrange Hair          -25.232     38.297  -0.659   0.5101    
data$HAIRPink Hair            -53.020     70.107  -0.756   0.4496    
data$HAIRPlatinum Blond Hair  -46.241     98.807  -0.468   0.6398    
data$HAIRPurple Hair          -44.652     36.670  -1.218   0.2235    
data$HAIRRed Hair             -14.654     11.580  -1.265   0.2058    

data$HAIRReddish Brown Hair      7.004    139.308   0.050   0.9599  
data$HAIRStrawberry Blond Hair  -5.602     33.472  -0.167   0.8671  
data$HAIRViolet Hair            -4.965     80.471  -0.062   0.9508  
data$HAIRWhite Hair             -8.137     13.100  -0.621   0.5346  
data$SEXGenderless Characters  -11.808     82.443  -0.143   0.8861  
data$SEXMale Characters         16.976      6.620   2.564   0.0104 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

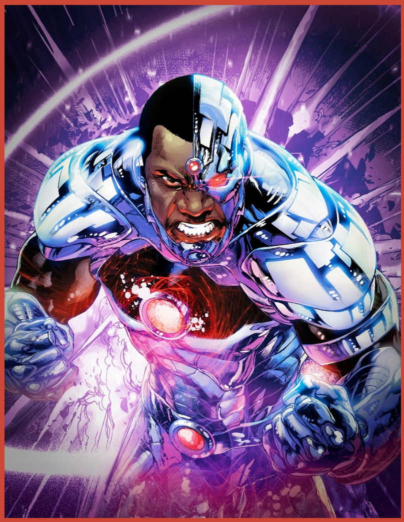
Residual standard error: 139 on 2069 degrees of freedom
  (72 observations deleted due to missingness)
Multiple R-squared:  0.05694,	Adjusted R-squared:  0.04007 
F-statistic: 3.376 on 37 and 2069 DF,  p-value: 4.673e-11
```

ANALYSIS OF LINEAR REGRESSION

The baseline / intercept for this model is a character with an unknown identity, with bad character, amber eyes, black hair, and is female. The intercept indicates that a character with these attributes would have approximately 23.377 appearances. Each attribute has a corresponding coefficient unless included in the intercept. In order to approximate how many appearances a character would have, one would simply multiply 1 to each coefficient corresponding to the attributes the character has and add each of those to the intercept. If the character does not have an attribute, a zero is multiplied by the attribute and added to the intercept. All attributes with a negative coefficient indicate a negative correlation with the number of appearances and all those with a positive correlation indicate a positive correlation. However, the only 3 attributes that are statistically significant are good characters, male characters and reformed criminals according to their t value (absolute value greater than 1.96) and p value (smaller than 0.05). All 3 have a positive correlation with appearances as they all correspond with positive coefficients.

What is super important to note about this linear regression is that it has a very small multiple r-squared at 0.05694 and a very small adjusted r-squared of 0.04007. R squared is a statistical measure of how close the data are to the fitted regression line. How much of the data can be explained by the model. With such a small r squared, this linear model does not predict appearances very well with the inputs of the categorical variables. If I were to proceed with this analysis, it would be a good idea to include interaction terms with this model. I could also add in the variable if he/she is dead or alive, a variable in the original data set that I did not use.

SECONDARY RESEARCH



I initially wanted to look at how the ethnicity of a comic book character affected the number of appearances he/she had in the comics, but couldn't find a dataset that addressed that information. Additionally, DC has a lot of characters who are racially ambiguous. I decided to do some outside research on the topic.

Up until a couple of years ago, DC faced a plethora of problems regarding diversity. The company had whitewashed characters during Black History Month. For example, Helena Bertinelli's version of the Huntress was portrayed as a woman of color but in subsequent comics, her skin color was inconsistent. There were images of her next to her white counterparts where there was no noticeable difference in skin tone. This has been a common theme in DC as also shown with Connor Hawke, the second Green Arrow. He first appeared to be a mix of Korean, white, and black heritage, but later became strictly white ethically with blond hair. Another thing to note is DC's tendency to have women playing characters complimentary to their male counterparts like how Wonder Woman played Superman's girlfriend even though she herself is a God and superhero in her own right.

Luckily DC is slowly but surely shifting the tide, becoming a whole lot more inclusive. Catwoman had a storyline revealing her bisexuality and Bombshells has gained significany traction, follows a reality where female superheroes protect the homefront during WW2. DC also relaunched Milestone Comics, a black universe of superheroes such as one of my favorites, Static Shock. There even is going to be a special by infamous screenwriter John Ridley on a Black Batman.

SOMETHING TO NOTE

A big challenge of this data set and my model is that it does not take into account that a comic book character may appear more frequently because he/she was introduced earlier. It would be important to take this into consideration if I were to further examine this data. This does however fit into the narrative that certain types of characters were introduced earlier because of the times. The first comic book characters were dominated by men due to the nature of society. Also, the three most popular superheros, Superman , Batman, and Wonder Woman were introduced very early on and had blue eyes, a eurocentric feature and were all white .

RECOMMENDATIONS

MORE FEMALE CHARACTERS

There needs to be more female characters, there are still way more male characters precentage wise

MORE CHARACTERS FROM DIFFERENT ETHINC BACKGROUNDS

There needs to be characters from all backgrounds. Comic books shouldn't be filled with only white characters. There should be higher presentation of minorities.

MORE CHARACTERS FROM THE LGBTQ COMMUNITY

There should be more characters like Catwoman who represents a part of the LGBTQ with her sexuality. There should be more genderless characters as well not just a couple.

REFERENCES

Diversity is making DC Comics great again

Data Set

DC Comics' Latest Event Showed The Necessity Of Diversity In Superhero Media

HUNTRESS AND DC'S PROBLEM WITH WHITEWASHING