Comparison of Marvel and DC Characters

Team Members: [Jizhou Cheng, Jiahao Li]

2024-12-18

Introduction

This report aims to compare Marvel and DC characters based on three key attributes: **eye color**, **alignment**, and **gender**. The purpose of this analysis is to identify patterns and differences between the two companies' character designs, providing insights into their creative strategies and diversity.

We used publicly available data sources, cleaned the raw data, and visualized the results through tables and charts. Statistical tests were performed to validate the observed differences.

Purpose

The purpose of this study was to analyze differences in the eye color, alignment, and gender distributions of Marvel and DC characters to identify patterns and trends in their character designs. This analysis aims to uncover:

- Differences in creative approaches between the two companies.
- Insights into character diversity and design trends.
- Key aspects that make each company's characters unique.

Data Description

The dataset contains the following columns:

- `page_id`: Unique ID for the character's page.
- `name`: Name of the character.
- `eye`: Eye color of the character.

- `align`: Alignment (Good, Neutral, Evil).
- `sex`: Gender of the character (Male, Female, etc.).
- `gsm`: Gender or sexual minority status.
- `alive`: Character status (alive or deceased).
- `company`: Data source (Marvel or DC).

Methods

Data Import and Cleaning

The raw data files were cleaned and processed as follows:

```r
# Load necessary packages
library(dplyr)
library(ggplot2)
library(knitr)

# Import raw data
marvel <- read.csv("~/Downloads/marvel-wikia-data.csv")
dc <- read.csv("~/Downloads/dc-wikia-data.csv")

# Standardize column names
names(marvel) <- tolower(trimws(names(marvel)))
names(dc) <- tolower(trimws(names(dc)))

# Combine datasets
combined_data <- bind_rows(
  marvel %>% mutate(company = "Marvel"),
  dc %>% mutate(company = "DC")
)

# Remove missing or blank values
cleaned_data <- combined_data %>%
  filter(!is.na(name) & name != "") %>%
  mutate(
    eye = tolower(trimws(eye)),
    align = tolower(trimws(align)),
    sex = tolower(trimws(sex))
  )
```

```r
# Save cleaned data to CSV
write.csv(cleaned_data, "~/Downloads/cleaned-comic-characters.csv", row.names = FALSE)
```

Variables of Interest

The three attributes analyzed were: - **Eye Color**: Distribution and comparison of eye colors across Marvel and DC characters. - **Alignment**: Analysis of alignments (Good, Neutral, Evil) for characters in each company. - **Gender**: Examination of gender distributions.

Results

Eye Color Distribution

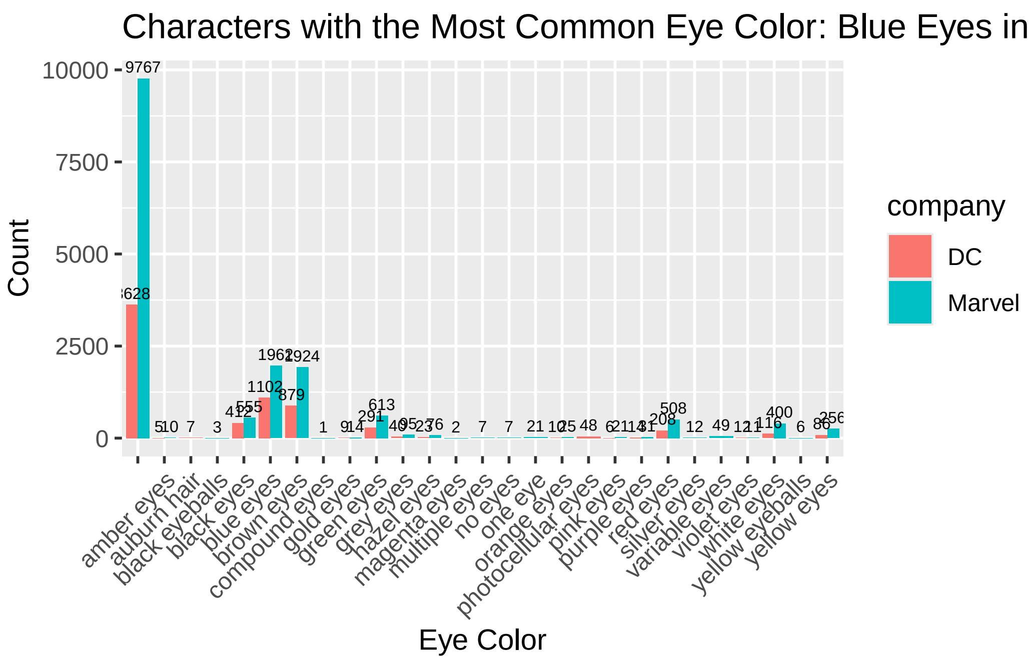
Table: Eye Color Distribution

Table 1: Eye Color Distribution by Company

company	eye	count
DC		3628
DC	amber eyes	5
DC	auburn hair	7
DC	black eyes	412
DC	blue eyes	1102
DC	brown eyes	879
DC	gold eyes	9
DC	green eyes	291
DC	grey eyes	40
DC	hazel eyes	23
DC	orange eyes	10
DC	photocellular eyes	48
DC	pink eyes	6
DC	purple eyes	14
DC	red eyes	208
DC	violet eyes	12
DC	white eyes	116
DC	yellow eyes	86
Marvel		9767
Marvel	amber eyes	10
Marvel	black eyeballs	3

company	eye	count
Marvel	black eyes	555
Marvel	blue eyes	1962
Marvel	brown eyes	1924
Marvel	compound eyes	1
Marvel	gold eyes	14
Marvel	green eyes	613
Marvel	grey eyes	95
Marvel	hazel eyes	76
Marvel	magenta eyes	2
Marvel	multiple eyes	7
Marvel	no eyes	7
Marvel	one eye	21
Marvel	orange eyes	25
Marvel	pink eyes	21
Marvel	purple eyes	31
Marvel	red eyes	508
Marvel	silver eyes	12
Marvel	variable eyes	49
Marvel	violet eyes	11
Marvel	white eyes	400
Marvel	yellow eyeballs	6
Marvel	yellow eyes	256

Chart: Characters with the Most Common Eye Color: Blue Eyes in Marvel and DC



Pearson's Chi-Square Test for Eye Color

Objective: To determine whether the eye color distributions between Marvel and DC characters differ significantly and whether the likelihood of eye color categories is the same for both companies.

-Null Hypothesis (H): The eye color distributions for Marvel and DC characters are the same. -Alternative Hypothesis (H): The eye color distributions for Marvel and DC characters are different. -Significance Level ( ): 0.05

```
    Pearson's Chi-squared test

data:  eye_contingency
X-squared = 387.13, df = 26, p-value < 2.2e-16


                    company
eye                            DC       Marvel
                     3969.2299759 9425.7700241
  amber eyes            4.4448264   10.5551736
```

```
auburn hair           2.0742523    4.9257477
black eyeballs        0.8889653    2.1110347
black eyes          286.5431420  680.4568580
blue eyes           907.9298728 2156.0701272
brown eyes          830.5898934 1972.4101066
compound eyes         0.2963218    0.7036782
gold eyes             6.8154005   16.1845995
green eyes          267.8748711  636.1251289
grey eyes            40.0034376   94.9965624
hazel eyes           29.3358542   69.6641458
magenta eyes          0.5926435    1.4073565
multiple eyes         2.0742523    4.9257477
no eyes               2.0742523    4.9257477
one eye               6.2227570   14.7772430
orange eyes          10.3712616   24.6287384
photocellular eyes   14.2234445   33.7765555
pink eyes             8.0006875   18.9993125
purple eyes          13.3344792   31.6655208
red eyes            212.1663802  503.8336198
silver eyes           3.5558611    8.4441389
variable eyes        14.5197662   34.4802338
violet eyes           6.8154005   16.1845995
white eyes          152.9020282  363.0979718
yellow eyeballs       1.7779306    4.2220694
yellow eyes         101.3420419  240.6579581
```

Process:

1.Assume the eye color distributions for Marvel and DC characters are the same (H). 2.Perform Pearson's Chi-squared test to analyze the distribution differences. 3.Evaluate the p-value from the test: -If p-value < 0.05, reject H , indicating a significant difference. -If p-value 0.05, do not reject H , indicating no significant difference.

Results:

- **Chi-Square Statistic**: 387.13
- **Degrees of Freedom**: 26
- **p-value**: $<$ 2.2e-16

Conclusion:

Since the p-value is extremely small ($p < 0.05$), we reject the null hypothesis. Therefore, we conclude that the eye color distributions between Marvel and DC characters differ significantly.

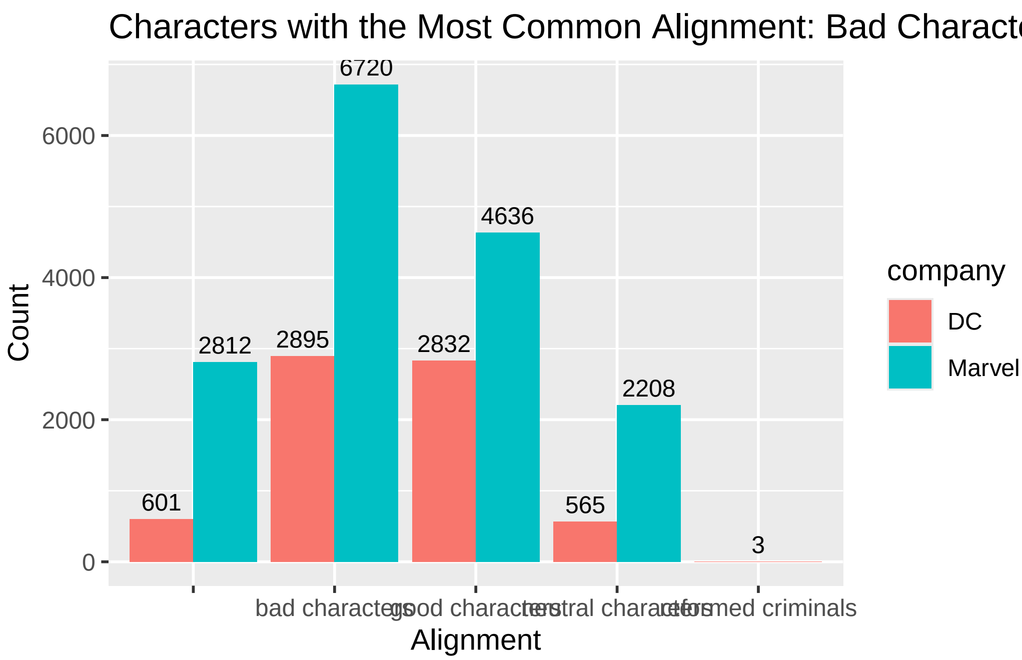
Alignment Distribution

Table: Alignment Distribution

Table 2: Alignment Distribution by Company

company	align	count
DC		601
DC	bad characters	2895
DC	good characters	2832
DC	neutral characters	565
DC	reformed criminals	3
Marvel		2812
Marvel	bad characters	6720
Marvel	good characters	4636
Marvel	neutral characters	2208

Chart: Characters with the Most Common Alignment: Bad Characters in Marvel and DC



Pearson's Chi-Square Test for Alignment

Objective: To determine whether the alignment distributions between Marvel and DC characters differ significantly and whether the likelihood of alignment categories is the same for both companies.

-Null Hypothesis (H): The alignment distributions for Marvel and DC characters are the same. -Alternative Hypothesis (H): The alignment distributions for Marvel and DC characters are different. -Significance Level ( ): 0.05

```
    Pearson's Chi-squared test

data:  align_contingency
X-squared = 604.86, df = 4, p-value < 2.2e-16


                    company
align                          DC      Marvel
                    1011.3461671 2401.653833
  bad characters     2849.1337229 6765.866277
```

```
good characters    2212.9309041 5255.069096
neutral characters  821.7002406 1951.299759
reformed criminals    0.8889653    2.111035
```

Process:

1.Assume the alignment distributions for Marvel and DC characters are the same (H). 2.Perform Pearson's Chi-squared test to analyze the distribution differences. 3.Evaluate the p-value from the test: -If p-value $<$ 0.05, reject H , indicating a significant difference. -If p-value 0.05, do not reject H , indicating no significant difference.

Results:

- **Chi-Square Statistic**: 604.86
- **Degrees of Freedom**: 4
- **p-value**: $<$ 2.2e-16

Conclusion:

Since the p-value is extremely small (p $<$ 0.05), we reject the null hypothesis. Therefore, we conclude that the alignment distributions between Marvel and DC characters differ significantly.

Gender Distribution

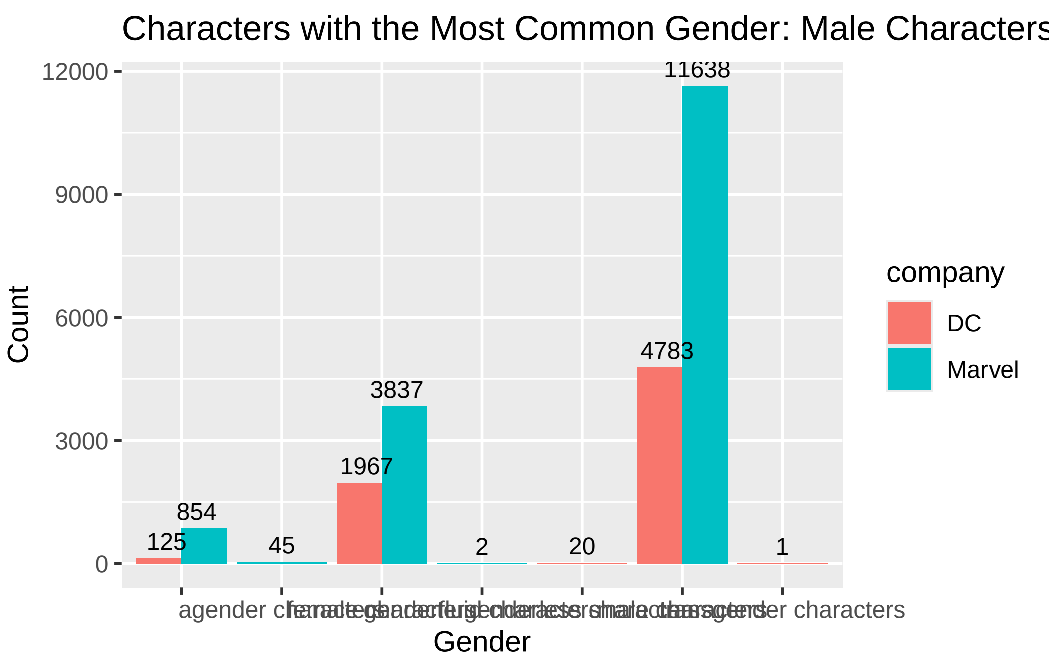
Table: Gender Distribution

Table 3: Gender Distribution by Company

company	sex	count
DC		125
DC	female characters	1967
DC	genderless characters	20
DC	male characters	4783
DC	transgender characters	1
Marvel		854
Marvel	agender characters	45

company	sex	count
Marvel	female characters	3837
Marvel	genderfluid characters	2
Marvel	male characters	11638

Chart: Characters with the Most Common Gender: Male Characters in Marvel and DC



Pearson's Chi-Square Test for Gender

Objective: To determine whether the gender distributions between Marvel and DC characters differ significantly and whether the likelihood of gender categories is the same for both companies. -Null Hypothesis (H): The gender distributions for Marvel and DC characters are the same. -Alternative Hypothesis (H): The gender distributions for Marvel and DC characters are different. -Significance Level ( ): 0.05

```
    Pearson's Chi-squared test

data:  gender_contingency
X-squared = 255.67, df = 6, p-value < 2.2e-16
```

```
                        company
sex                                DC       Marvel
                          290.0990031 6.889010e+02
  agender characters       13.3344792 3.166552e+01
  female characters      1719.8514954 4.084149e+03
  genderfluid characters    0.5926435 1.407356e+00
  genderless characters     5.9264352 1.407356e+01
  male characters        4865.8996219 1.155510e+04
  transgender characters    0.2963218 7.036782e-01
```

Process:

1.Assume the gender distributions for Marvel and DC characters are the same (H). 2.Perform Pearson's Chi-squared test to analyze the distribution differences. 3.Evaluate the p-value from the test: -If p-value < 0.05, reject H , indicating a significant difference. -If p-value 0.05, do not reject H , indicating no significant difference.

Results:

- **Chi-Square Statistic**: 255.67
- **Degrees of Freedom**: 6
- **p-value**: < 2.2e-16

Conclusion:

Since the p-value is extremely small ($p < 0.05$), we reject the null hypothesis. Therefore, we conclude that the gender distributions between Marvel and DC characters differ significantly.

Summary and Conclusions

Key Observations

1. **Eye Color**:
    - Marvel characters exhibit a wider variety of eye colors, including gold and red.
    - DC characters primarily feature traditional eye colors like blue and brown.
2. **Alignment**:
    - Marvel has more neutral characters compared to DC.
    - DC focuses on clear distinctions between good and bad characters.

3. **Gender**:

    - Both companies are male-dominated.
    - Marvel features slightly more female and non-traditional gender characters.

Conclusion

Marvel demonstrates a greater emphasis on diversity in eye color, alignment, and gender, while DC tends to focus on traditional traits and roles. These differences reflect distinct creative approaches and target audiences.

References

- Data source: jayb.fivethirtyeight. FiveThirtyEight Comic Characters.
- Marvel data: By Fandom team. Marvel Wikia
- DC data: By Fandom team. DC Wikia

Appendix

```
# Load necessary packages
library(dplyr)
library(ggplot2)
library(knitr)

# Import raw data
marvel <- read.csv("~/Downloads/marvel-wikia-data.csv")
dc <- read.csv("~/Downloads/dc-wikia-data.csv")

# Standardize column names
names(marvel) <- tolower(trimws(names(marvel)))
names(dc) <- tolower(trimws(names(dc)))

# Combine datasets
combined_data <- bind_rows(
  marvel %>% mutate(company = "Marvel"),
  dc %>% mutate(company = "DC")
)

# Remove missing or blank values
```

```
cleaned_data <- combined_data %>%
  filter(!is.na(name) & name != "") %>%
  mutate(
    eye = tolower(trimws(eye)),
    align = tolower(trimws(align)),
    sex = tolower(trimws(sex))
  )

# Save cleaned data to CSV
write.csv(cleaned_data, "~/Downloads/cleaned-comic-characters.csv", row.names = FALSE)
eye_distribution <- cleaned_data %>%
  group_by(company, eye) %>%
  summarise(count = n(), .groups = "drop")

# Table
knitr::kable(eye_distribution, caption = "Eye Color Distribution by Company")
# Plot eye color with labels
ggplot(eye_distribution, aes(x = eye, y = count, fill = company)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label = count), position = position_dodge(width = 0.8), vjust = -0.5, size =
  labs(title = "Characters with the Most Common Eye Color: Blue Eyes in Marvel and DC", x = "
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
# Contingency table for eye color and company
eye_contingency <- xtabs(~ eye + company, data = cleaned_data)

# Perform chi-square test
eye_test <- chisq.test(eye_contingency)

# Print results
eye_test

# Check expected values
eye_test$expected
align_distribution <- cleaned_data %>%
  group_by(company, align) %>%
  summarise(count = n(), .groups = "drop")

# Table
knitr::kable(align_distribution, caption = "Alignment Distribution by Company")
# Plot alignment with labels
ggplot(align_distribution, aes(x = align, y = count, fill = company)) +
  geom_bar(stat = "identity", position = "dodge") +
```

```r
  geom_text(aes(label = count), position = position_dodge(width = 0.9), vjust = -0.5, size =
  labs(title = "Characters with the Most Common Alignment: Bad Characters in Marvel and DC",
# Contingency table for alignment and company
align_contingency <- xtabs(~ align + company, data = cleaned_data)

# Perform chi-square test
align_test <- chisq.test(align_contingency)

# Print results
align_test

# Check expected values
align_test$expected
gender_distribution <- cleaned_data %>%
  group_by(company, sex) %>%
  summarise(count = n(), .groups = "drop")

# Table
knitr::kable(gender_distribution, caption = "Gender Distribution by Company")
# Plot gender distribution with labels
ggplot(gender_distribution, aes(x = sex, y = count, fill = company)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label = count), position = position_dodge(width = 0.6), vjust = -0.5, size =
  labs(title = "Characters with the Most Common Gender: Male Characters in Marvel and DC", x
# Contingency table for gender and company
gender_contingency <- xtabs(~ sex + company, data = cleaned_data)

# Perform chi-square test
gender_test <- chisq.test(gender_contingency)

# Print results
gender_test

# Check expected values
gender_test$expected
```