EDA of Categorical Data

Exploring categorical data

- creating graphical and numerical summaries of two categorical variables using the comics dataset

Explore data

In [35]:
```
# IMport relevant libraries
library(ggplot2)
library(dplyr)

options(scipen=999)
```

In [2]:
```
# define the filename
filename <- "comics.csv"

# load the CSV file from the local directory
comics <- read.csv(filename)

# preview the first 5 rows
head(comics)
```

name	id	align	eye	hair	gender	gsm	alive	appearances	first_appear	publisher
Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Brown Hair	Male	NA	Living Characters	4043	Aug-62	marvel
Captain America (Steven Rogers)	Public	Good	Blue Eyes	White Hair	Male	NA	Living Characters	3360	Mar-41	marvel
Wolverine (James \"Logan\" Howlett)	Public	Neutral	Blue Eyes	Black Hair	Male	NA	Living Characters	3061	Oct-74	marvel
Iron Man (Anthony \"Tony\" Stark)	Public	Good	Blue Eyes	Black Hair	Male	NA	Living Characters	2961	Mar-63	marvel
Thor (Thor Odinson)	No Dual	Good	Blue Eyes	Blond Hair	Male	NA	Living Characters	2258	Nov-50	marvel
Benjamin Grimm (Earth-616)	Public	Good	Blue Eyes	No Hair	Male	NA	Living Characters	2255	Nov-61	marvel

In [3]:
```
# Data on all commics created by DC and Marvel
comics
```

name	id	align	eye	hair	gender	gsm	alive	appearances	first_appear	publisher
Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Brown Hair	Male	NA	Living Characters	4043	Aug-62	marvel
Captain America (Steven Rogers)	Public	Good	Blue Eyes	White Hair	Male	NA	Living Characters	3360	Mar-41	marvel
Wolverine (James \"Logan\" Howlett)	Public	Neutral	Blue Eyes	Black Hair	Male	NA	Living Characters	3061	Oct-74	marvel
Iron Man (Anthony \"Tony\" Stark)	Public	Good	Blue Eyes	Black Hair	Male	NA	Living Characters	2961	Mar-63	marvel
Thor (Thor Odinson)	No Dual	Good	Blue Eyes	Blond Hair	Male	NA	Living Characters	2258	Nov-50	marvel
Benjamin Grimm (Earth-616)	Public	Good	Blue Eyes	No Hair	Male	NA	Living Characters	2255	Nov-61	marvel
Reed Richards (Earth-616)	Public	Good	Brown Eyes	Brown Hair	Male	NA	Living Characters	2072	Nov-61	marvel
Hulk (Robert Bruce Banner)	Public	Good	Brown Eyes	Brown Hair	Male	NA	Living Characters	2017	May-62	marvel
Scott Summers (Earth-616)	Public	Neutral	Brown Eyes	Brown Hair	Male	NA	Living Characters	1955	Sep-63	marvel
Jonathan Storm (Earth-616)	Public	Good	Blue Eyes	Blond Hair	Male	NA	Living Characters	1934	Nov-61	marvel
Henry McCoy (Earth-616)	Public	Good	Blue Eyes	Blue Hair	Male	NA	Living Characters	1825	Sep-63	marvel
Susan Storm (Earth-616)	Public	Good	Blue Eyes	Blond Hair	Female	NA	Living Characters	1713	Nov-61	marvel
Namor McKenzie (Earth-616)	No Dual	Neutral	Green Eyes	Black Hair	Male	NA	Living Characters	1528	NA	marvel
Ororo Munroe (Earth-616)	Public	Good	Blue Eyes	White Hair	Female	NA	Living Characters	1512	May-75	marvel
Clinton Barton (Earth-616)	Public	Good	Blue Eyes	Blond Hair	Male	NA	Living Characters	1394	Sep-64	marvel
Matthew Murdock (Earth-616)	Public	Good	Blue Eyes	Red Hair	Male	NA	Living Characters	1338	Apr-64	marvel
Stephen Strange (Earth-616)	Public	Good	Grey Eyes	Black Hair	Male	NA	Living Characters	1307	Jul-63	marvel
Mary Jane Watson (Earth-616)	No Dual	Good	Green Eyes	Red Hair	Female	NA	Living Characters	1304	Jun-65	marvel
John Jonah Jameson (Earth-616)	No Dual	Neutral	Blue Eyes	Black Hair	Male	NA	Living Characters	1266	Mar-63	marvel
Robert Drake (Earth-616)	Secret	Good	Brown Eyes	Brown Hair	Male	NA	Living Characters	1265	Sep-63	marvel
Henry Pym (Earth-616)	Public	Good	Blue Eyes	Blond Hair	Male	NA	Living Characters	1237	Jan-62	marvel
Charles Xavier (Earth-616)	Public	Good	Blue Eyes	Bald	Male	NA	Deceased Characters	1233	Sep-63	marvel
Warren Worthington III (Earth-616)	Public	Good	Blue Eyes	Blond Hair	Male	NA	Living Characters	1230	Sep-63	marvel
Piotr Rasputin (Earth-616)	Secret	Good	Blue Eyes	Black Hair	Male	NA	Living Characters	1162	May-75	marvel
Wanda Maximoff (Earth-616)	Public	Good	Green Eyes	Brown Hair	Female	NA	Living Characters	1161	Mar-64	marvel
Nicholas Fury (Earth-616)	No Dual	Neutral	Brown Eyes	Brown Hair	Male	NA	Living Characters	1137	May-63	marvel
Janet van Dyne (Earth-616)	Public	Good	Blue Eyes	Auburn Hair	Female	NA	Living Characters	1120	Jun-63	marvel
Jean Grey (Earth-616)	Public	Good	Green Eyes	Red Hair	Female	NA	Deceased Characters	1107	Sep-63	marvel
Natalia Romanova (Earth-616)	Public	Good	Green Eyes	Red Hair	Female	Bisexual Characters	Living Characters	1050	Apr-64	marvel
Kurt Wagner (Earth-616)	Secret	Good	Yellow Eyes	Blue Hair	Male	NA	Living Characters	1047	May-75	marvel
...	...	...	...	...	...	...	...	...	...	...
Baron Tyrano (New Earth)	Secret	Bad	Blue Eyes	NA	Male	NA	Living Characters	NA	1967, July	dc
Brains (New Earth)	NA	Good	NA	NA	Male	NA	Living Characters	NA	1967, April	dc
Cracker (New Earth)	NA	Good	NA	NA	Male	NA	Living Characters	NA	1967, June	dc
Hard Head (New Earth)	NA	Good	NA	Black Hair	Male	NA	Living Characters	NA	1967, April	dc
Zig-Zag (New Earth)	NA	Good	NA	NA	Male	NA	Living Characters	NA	1967, June	dc
Dragonfly (New Earth)	Secret	Bad	NA	Black Hair	Female	NA	Living Characters	NA	1966, June	dc
Carl Bradford (New Earth)	NA	Bad	NA	NA	Male	NA	Living Characters	NA	1966, January	dc
Donna Troy (New Earth)	Public	Good	Blue Eyes	Black Hair	Female	NA	Living Characters	NA	1965, July	dc
Bartholomew Magan (New Earth)	NA	Bad	NA	NA	Male	NA	Living Characters	NA	1963, September	dc
James Moon (New Earth)	NA	NA	NA	NA	Male	NA	Living Characters	NA	1962, March	dc
Flash (Wally West)	Secret	Good	Green Eyes	Red Hair	Male	NA	Living Characters	NA	1960, January	dc
J'onn J'onzz (New Earth)	Public	Good	Red Eyes	NA	Male	NA	Living Characters	NA	1955, November	dc
Dorothea Tane (New Earth)	NA	NA	NA	Blond Hair	Female	NA	Living Characters	NA	1948, August	dc
Dmane (Earth-Two)	NA	Bad	Blue Eyes	NA	Male	NA	Living Characters	NA	1946, April	dc
Maximillian O'Leary (New Earth)	Public	Good	NA	Black Hair	Male	NA	Living Characters	NA	1946, January	dc
Doris Zuel (New Earth)	Secret	Bad	Green Eyes	Red Hair	Female	NA	Living Characters	NA	1944, June	dc
Doris Lee (New Earth)	Public	NA	Brown Eyes	Brown Hair	Female	NA	Deceased Characters	NA	1941, April	dc
Patrick O'Brian (New Earth)	Secret	Good	Blue Eyes	Black Hair	Male	NA	Living Characters	NA	1941, August	dc
Basil Karlo (New Earth)	Secret	Bad	Black Eyes	Black Hair	Male	NA	Living Characters	NA	1940, June	dc
Catwoman (Selina Kyle)	Secret	Neutral	Green Eyes	Black Hair	Female	NA	Living Characters	NA	1940, June	dc
Bedivere (New Earth)	NA	NA	NA	NA	Male	NA	Living Characters	NA	1936, February	dc
Herbert Hoover (New Earth)	Public	Good	NA	NA	Male	NA	Living Characters	NA	NA	dc
William Howard Taft (New Earth)	Public	Good	NA	NA	Male	NA	Living Characters	NA	NA	dc
Frank Fitzsimmons (New Earth)	Public	Good	NA	Grey Hair	Male	NA	Living Characters	NA	NA	dc
James Garfield (New Earth)	Public	Good	NA	NA	Male	NA	Living Characters	NA	NA	dc
Nadine West (New Earth)	Public	Good	NA	NA	Female	NA	Living Characters	NA	NA	dc
Warren Harding (New Earth)	Public	Good	NA	NA	Male	NA	Living Characters	NA	NA	dc
William Harrison (New Earth)	Public	Good	NA	NA	Male	NA	Living Characters	NA	NA	dc
William McKinley (New Earth)	Public	Good	NA	NA	Male	NA	Living Characters	NA	NA	dc
Mookie (New Earth)	Public	Bad	Blue Eyes	Blond Hair	Male	NA	Living Characters	NA	NA	dc

Working with factors

In [4]:
```
# Getting the levels of a specific variable
levels(comics$align)
```

1. 'Bad'
2. 'Good'
3. 'Neutral'
4. 'Reformed Criminals'

In [5]:
```
# Getting the levels of a specific variable
levels(comics$id)
```

1. 'No Dual'
2. 'Public'
3. 'Secret'
4. 'Unknown'

In [26]:
```
# Create a 2-way contingency table
table(comics$id, comics$align)
```

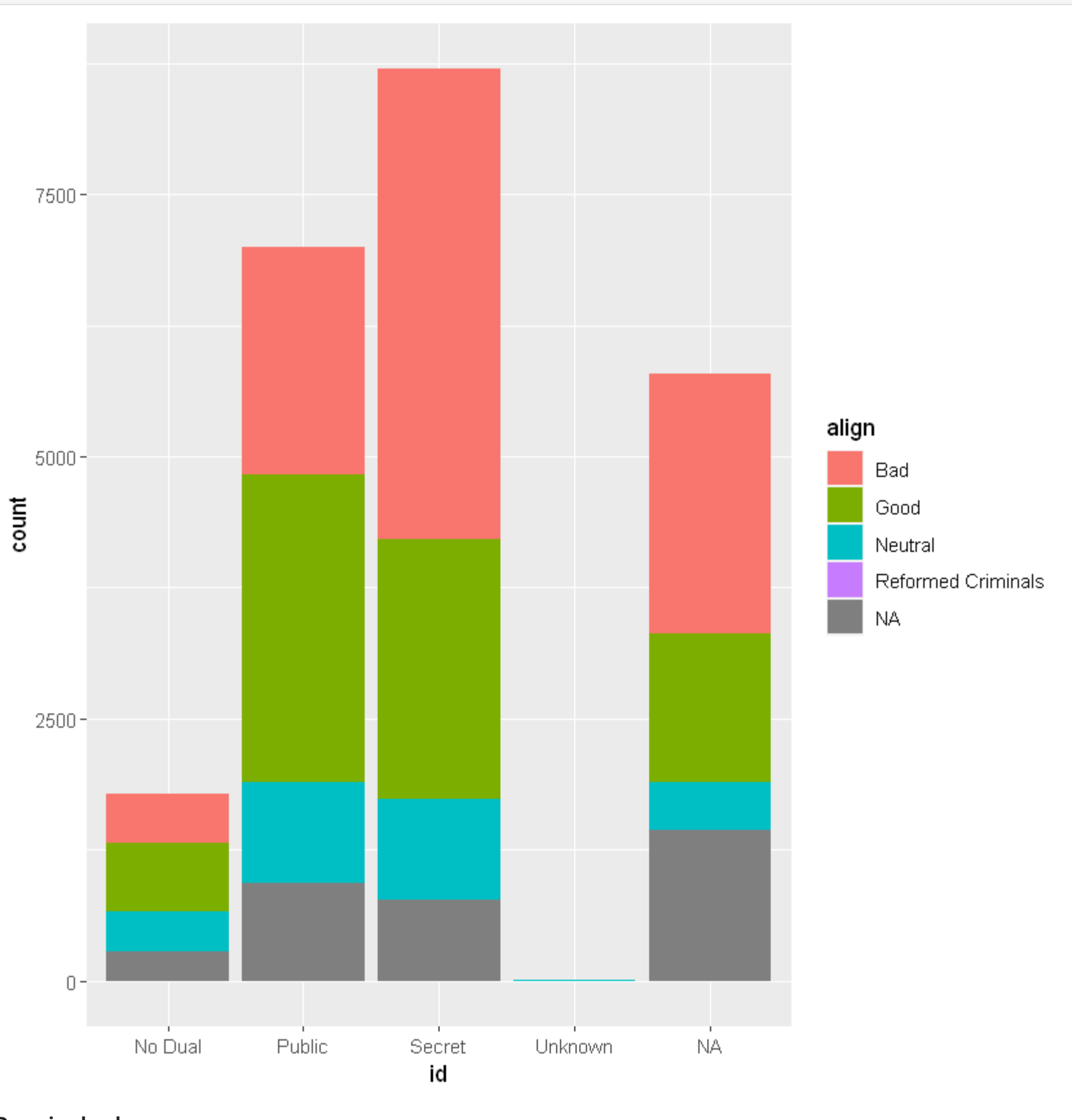
```
         Bad Good Neutral Reformed Criminals
 No Dual  474  647     390                  0
 Public  2172 2930     965                  1
 Secret  4493 2475     959                  1
 Unknown    7    0       2                  0
```

Insight from above table Most common identity are bad characters with hidden identity

In [27]:
```
# Save the contigency as a table
tab_cnt <- table(comics$id, comics$align)
tab_cnt
```

```
         Bad Good Neutral Reformed Criminals
 No Dual  474  647     390                  0
 Public  2172 2930     965                  1
 Secret  4493 2475     959                  1
 Unknown    7    0       2                  0
```

In [28]:
```
# Translate the table into graphs
ggplot(comics, aes(x = id, fill = align)) +
geom_bar()
```



Dropping levels

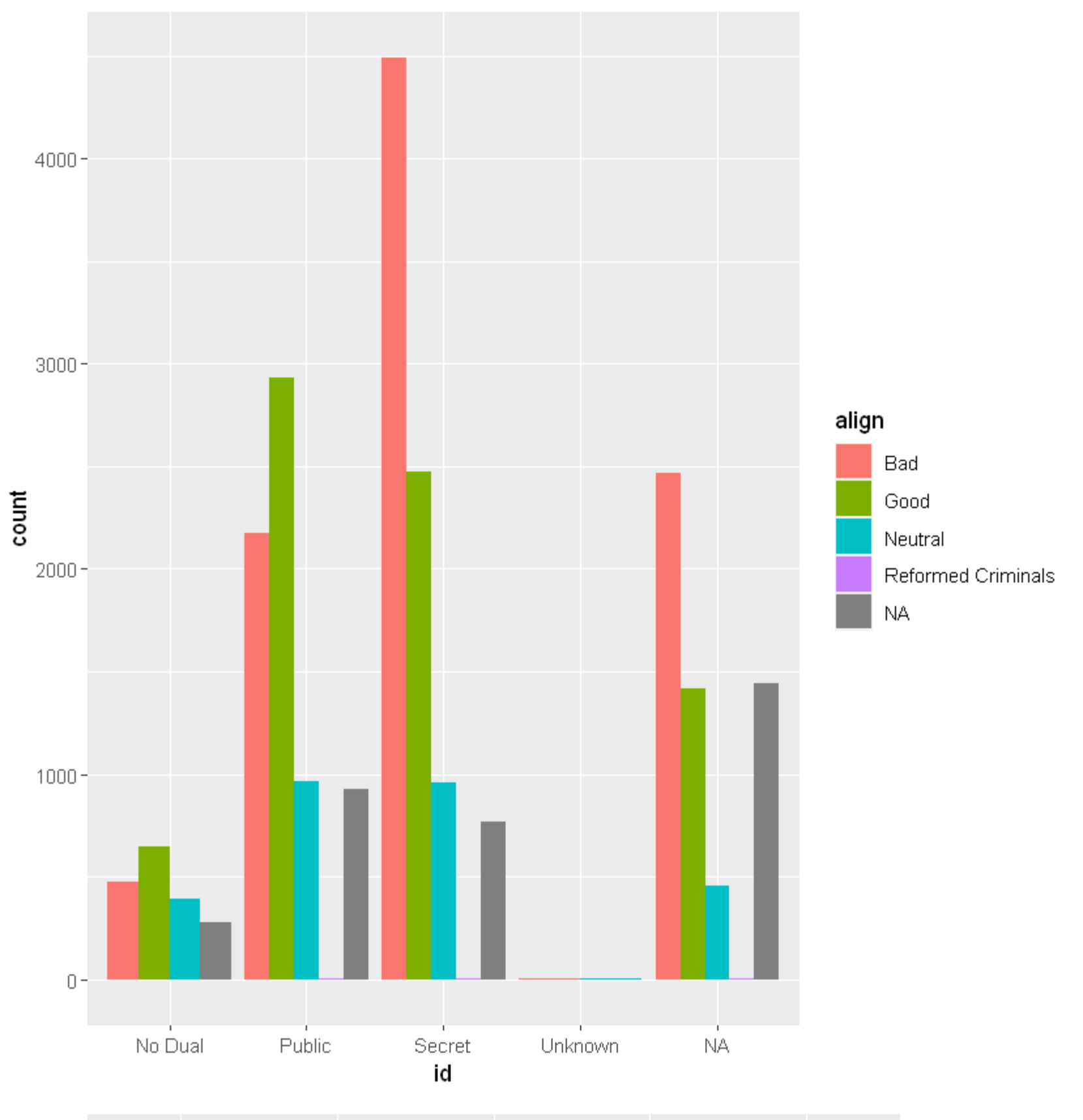
In [29]:
```
# Remove align level
# Use filter() to filter out all rows of comics with that level
# then drop the unused level with droplevels()
# Save the simplified dataset as comics_filtered
comics_filtered <- comics %>%
  filter(align != "Reformed Criminals") %>%
  droplevels()

# See the result
head(comics_filtered)

# Create a 2-way contingency table
table(comics_filtered$id, comics_filtered$align)
```

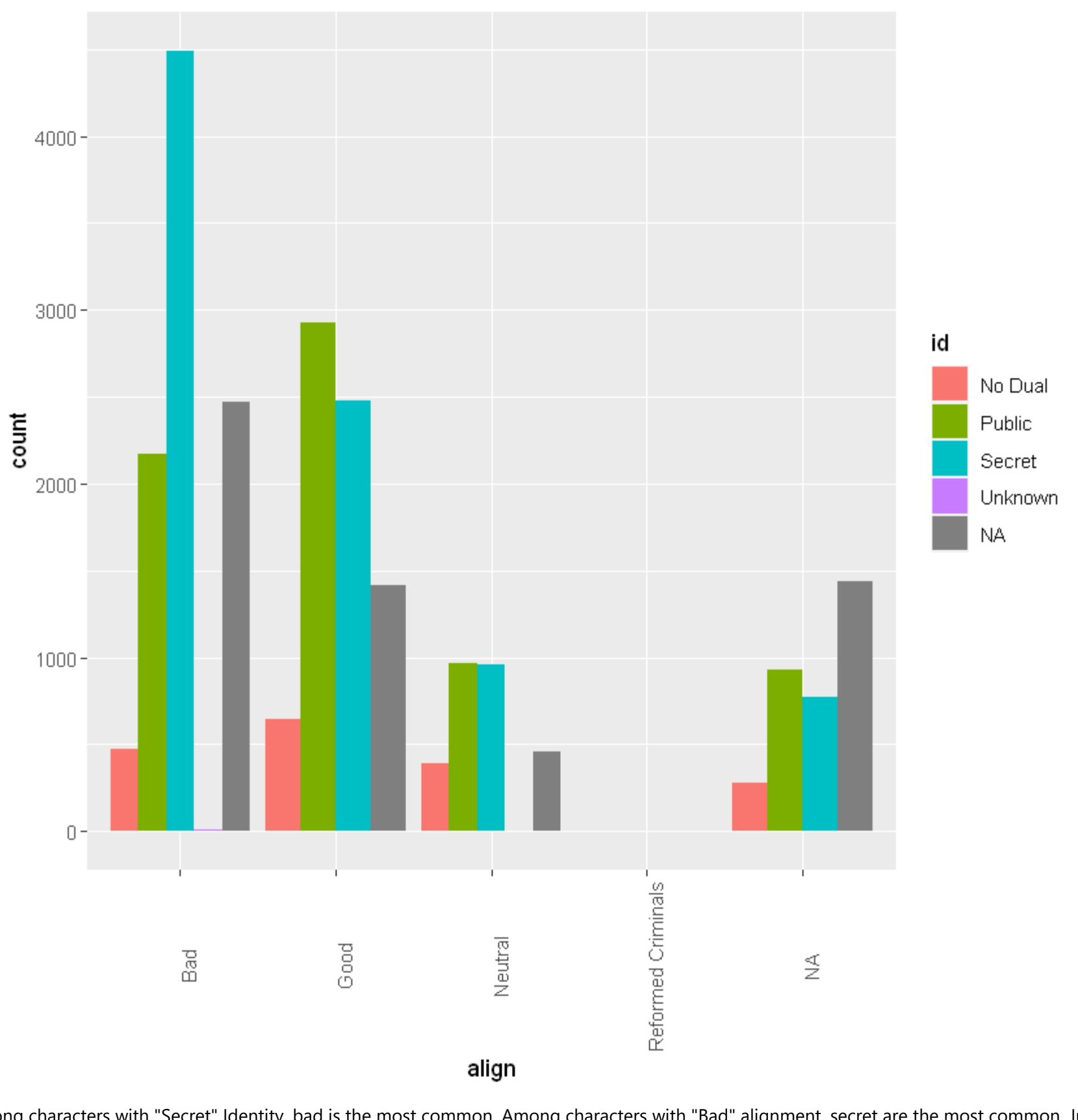
name	id	align	eye	hair	gender	gsm	alive	appearances	first_appear	publisher
Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Brown Hair	Male	NA	Living Characters	4043	Aug-62	marvel
Captain America (Steven Rogers)	Public	Good	Blue Eyes	White Hair	Male	NA	Living Characters	3360	Mar-41	marvel
Wolverine (James \"Logan\" Howlett)	Public	Neutral	Blue Eyes	Black Hair	Male	NA	Living Characters	3061	Oct-74	marvel
Iron Man (Anthony \"Tony\" Stark)	Public	Good	Blue Eyes	Black Hair	Male	NA	Living Characters	2961	Mar-63	marvel
Thor (Thor Odinson)	No Dual	Good	Blue Eyes	Blond Hair	Male	NA	Living Characters	2258	Nov-50	marvel
Benjamin Grimm (Earth-616)	Public	Good	Blue Eyes	No Hair	Male	NA	Living Characters	2255	Nov-61	marvel

```
         Bad Good Neutral
 No Dual  474  647     390
 Public  2172 2930     965
 Secret  4493 2475     959
 Unknown    7    0       2
```

Side-by-side barcharts

In [30]:
```
# Create side-by-side barchart of gender by alignment
ggplot(comics, aes(x = id, fill = align)) +
  geom_bar(position = "dodge")

# Create side-by-side barchart of alignment by gender
ggplot(comics, aes(x = align, fill = id)) +
  geom_bar(position = "dodge") +
  theme(axis.text.x = element_text(angle = 90))
```





Insight Among characters with "Secret" Identity, bad is the most common. Among characters with "Bad" alignment, secret are the most common. In general, there is an association between identity and alignment.

Counts vs Proportion

From counts to proportions

In [19]:
```
# Counts of the data
options(scipen = 999, digits = 3) # Simplify display format
tab_cnt <- table(comics$id, comics$align)
tab_cnt
```

```
         Bad Good Neutral Reformed Criminals
 No Dual  474  647     390                  0
 Public  2172 2930     965                  1
 Secret  4493 2475     959                  1
 Unknown    7    0       2                  0
```

In [20]:
```
# Proportions of data based on grand total
prop.table(tab_cnt)
```

```
               Bad      Good   Neutral Reformed Criminals
 No Dual 0.0305491 0.0416989 0.0251353          0.0000000
 Public  0.1399845 0.1888373 0.0621939          0.0000644
 Secret  0.2895721 0.1595128 0.0618072          0.0000644
 Unknown 0.0004511 0.0000000 0.0001289          0.0000000
```

In [21]:
```
# Sum of all these proportions
sum(prop.table(tab_cnt))
```

1

Conditional proportions

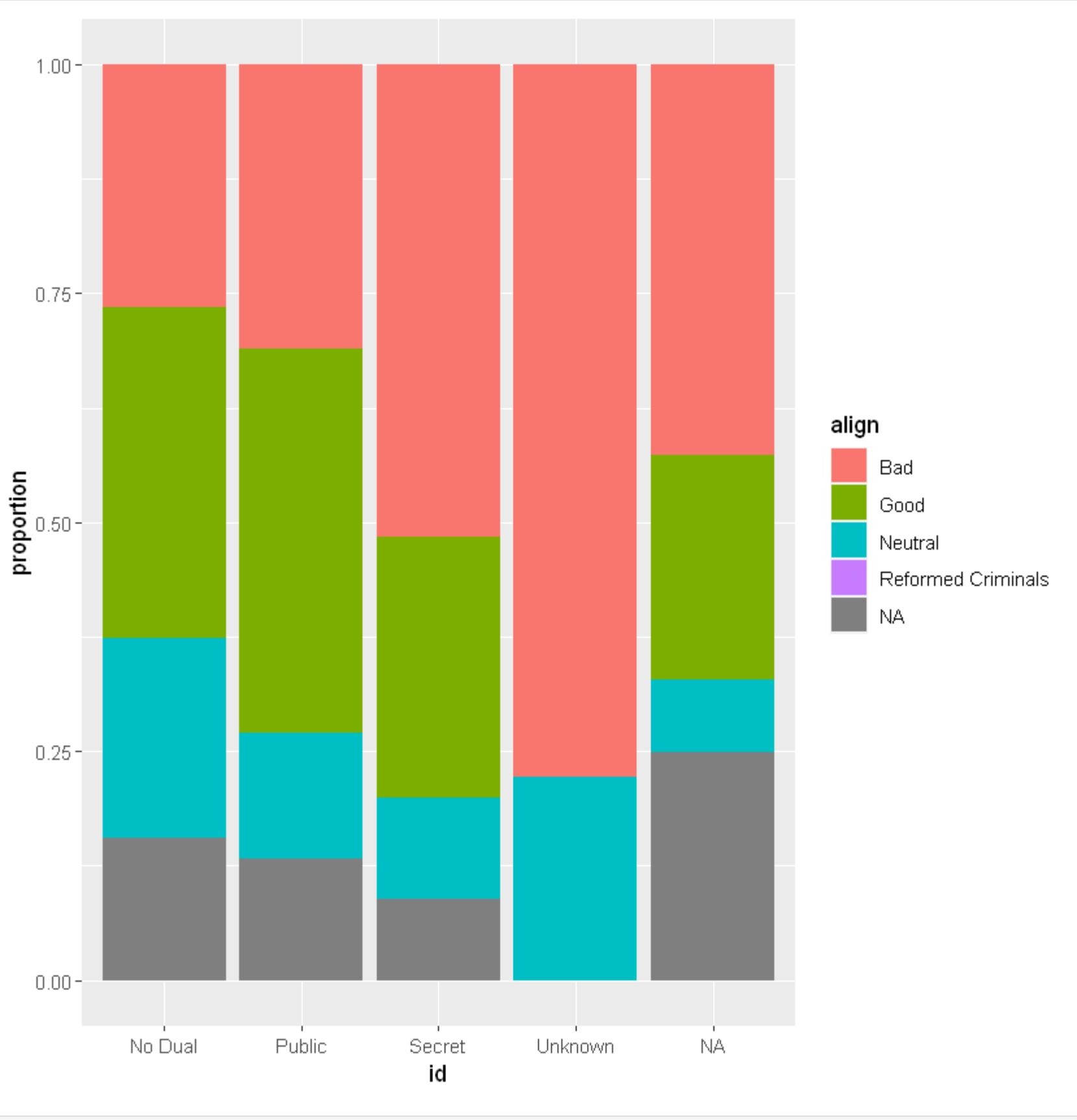
In [22]:
```
# Proportions of data based on row total
# Condition on Identity
prop.table(tab_cnt, 1)
```

```
              Bad     Good  Neutral Reformed Criminals
 No Dual 0.313700 0.428193 0.258107           0.000000
 Public  0.357943 0.482861 0.159031           0.000165
 Secret  0.566726 0.312185 0.120964           0.000126
 Unknown 0.777778 0.000000 0.222222           0.000000
```

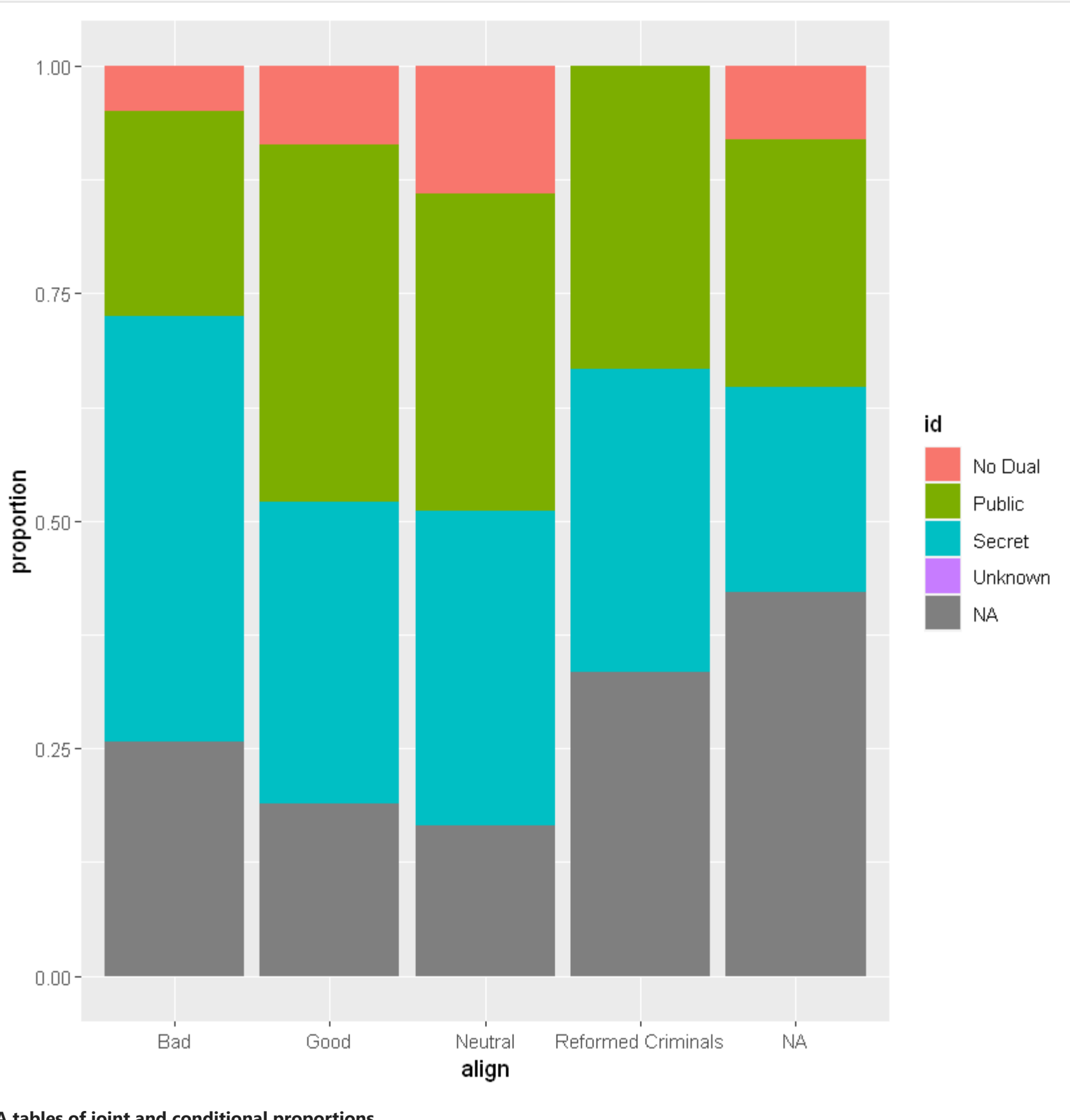
In [23]:
```
# Proportions of data based on column total
# Condition on alignment
prop.table(tab_cnt, 2)
```

```
              Bad     Good  Neutral Reformed Criminals
 No Dual 0.066331 0.106907 0.168394           0.000000
 Public  0.303946 0.484137 0.416667           0.500000
 Secret  0.628743 0.408956 0.414076           0.500000
 Unknown 0.000980 0.000000 0.000864           0.000000
```

In [24]:
```
# Bar chart based on proportions
ggplot(comics, aes(x = id, fill = align)) +
geom_bar(position = "fill") +
ylab("proportion")
```



In [25]:
```
ggplot(comics, aes(x = align, fill = id)) +
geom_bar(position = "fill") +
ylab("proportion")
```



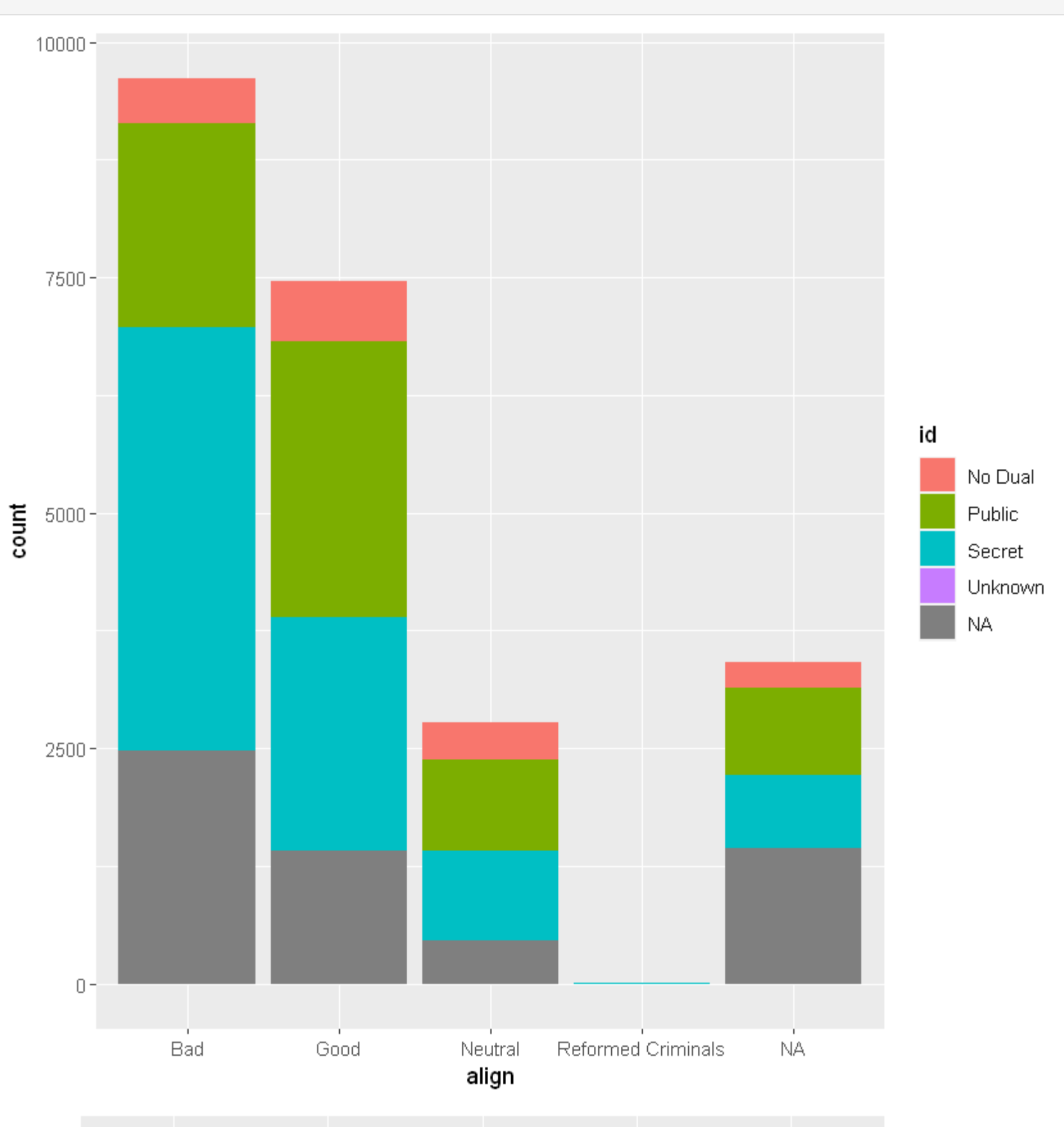
A tables of joint and conditional proportions

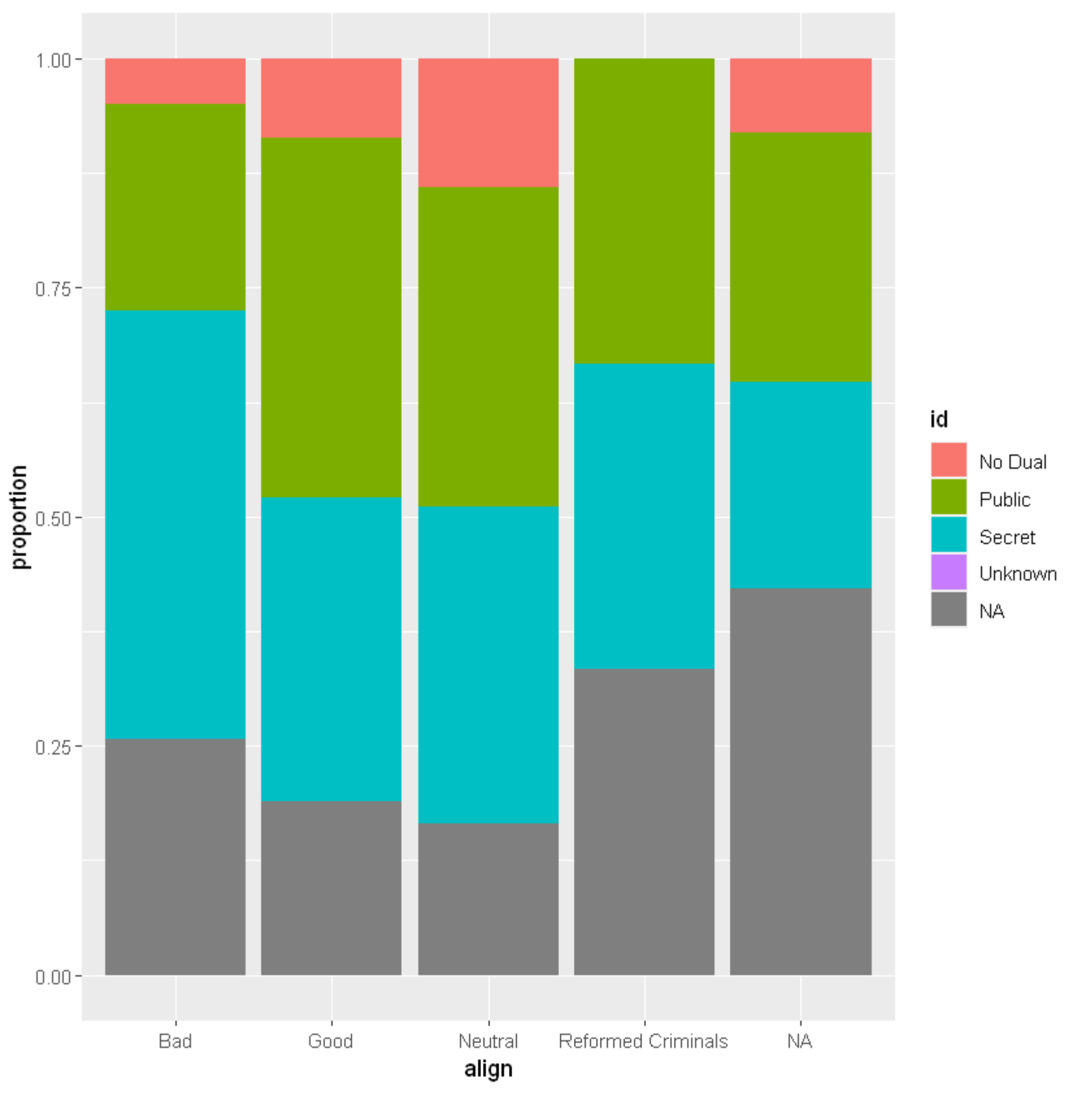
In [42]:
```
#  tables of joint and conditional proportions, respectively
tab <- table(comics$align, comics$id)

# Print fewer digits
options(scipen = 999, digits = 3)

# Joint proportions
prop.table(tab)

# Conditional on row
prop.table(tab, 1)

# Conditional on columns
prop.table(tab, 2)
```

```
                     No Dual    Public    Secret   Unknown
 Bad               0.0305491 0.1399845 0.2895721 0.0004511
 Good              0.0416989 0.1888373 0.1595128 0.0000000
 Neutral           0.0251353 0.0621939 0.0618072 0.0001289
 Reformed Criminals 0.0000000 0.0000644 0.0000644 0.0000000

                    No Dual   Public   Secret  Unknown
 Bad               0.066331 0.303946 0.628743 0.000980
 Good              0.106907 0.484137 0.408956 0.000000
 Neutral           0.168394 0.416667 0.414076 0.000864
 Reformed Criminals 0.000000 0.500000 0.500000 0.000000

                    No Dual   Public   Secret  Unknown
 Bad               0.313700 0.357943 0.566726 0.777778
 Good              0.428193 0.482861 0.312185 0.000000
 Neutral           0.258107 0.159031 0.120964 0.222222
 Reformed Criminals 0.000000 0.000165 0.000126 0.000000
```

Side-by Side Plot: Counts vs. proportions

In [39]:
```
# Plot of id by align
ggplot(comics, aes(x = align, fill = id)) +
  geom_bar()

# Plot proportion of gender, conditional on align
ggplot(comics, aes(x = align, fill = id)) +
  geom_bar(position = "fill") +
  ylab("proportion")
```





Distribution of one variable

In [48]:
```
# Drop irrelevant levels
comics_filtered <- comics %>%
  filter(align != "Reformed Criminals") %>%
  filter(align != "NA") %>%
  droplevels()

# See the result
head(comics_filtered)

# Create a 2-way contingency table
table(comics_filtered$id, comics_filtered$align)
```

name	id	align	eye	hair	gender	gsm	alive	appearances	first_appear	publisher
Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Brown Hair	Male	NA	Living Characters	4043	Aug-62	marvel
Captain America (Steven Rogers)	Public	Good	Blue Eyes	White Hair	Male	NA	Living Characters	3360	Mar-41	marvel
Wolverine (James \"Logan\" Howlett)	Public	Neutral	Blue Eyes	Black Hair	Male	NA	Living Characters	3061	Oct-74	marvel
Iron Man (Anthony \"Tony\" Stark)	Public	Good	Blue Eyes	Black Hair	Male	NA	Living Characters	2961	Mar-63	marvel
Thor (Thor Odinson)	No Dual	Good	Blue Eyes	Blond Hair	Male	NA	Living Characters	2258	Nov-50	marvel
Benjamin Grimm (Earth-616)	Public	Good	Blue Eyes	No Hair	Male	NA	Living Characters	2255	Nov-61	marvel

```
         Bad Good Neutral
 No Dual  474  647     390
 Public  2172 2930     965
 Secret  4493 2475     959
 Unknown    7    0       2
```

In [49]:
```
# Get the distribution of the variable id
table(comics_filtered$id)
```

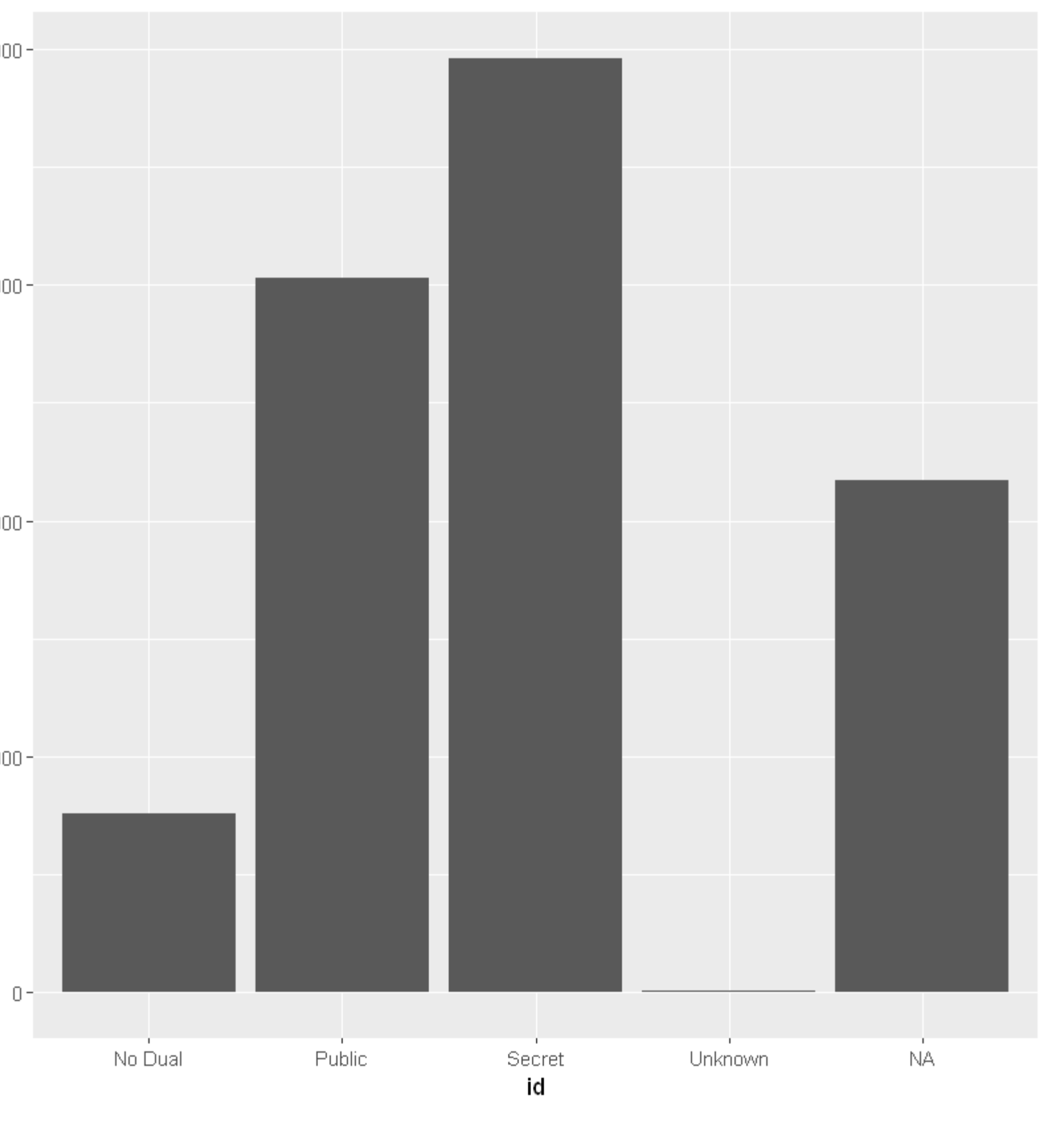
```
No Dual  Public  Secret Unknown 
   1511    6067    7927       9 
```

In [50]:
```
# Store the contingency table as a proper table
tab_cnt <- table(comics_filtered$id, comics_filtered$align)
tab_cnt
```

```
         Bad Good Neutral
 No Dual  474  647     390
 Public  2172 2930     965
 Secret  4493 2475     959
 Unknown    7    0       2
```

Graphs

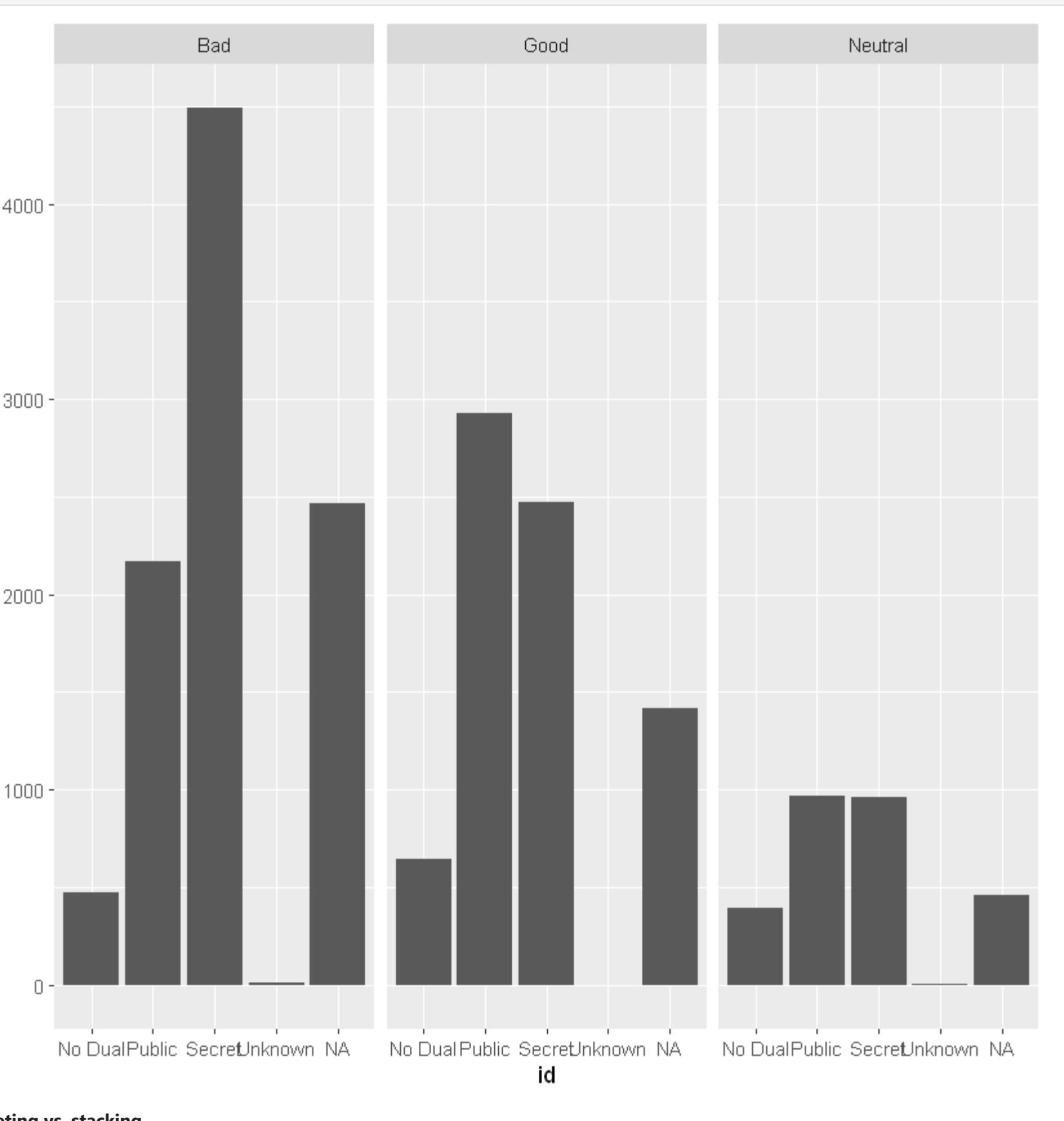
In [51]:
```
# Simple barchart
ggplot(comics_filtered, aes(x = id)) +
geom_bar()
```



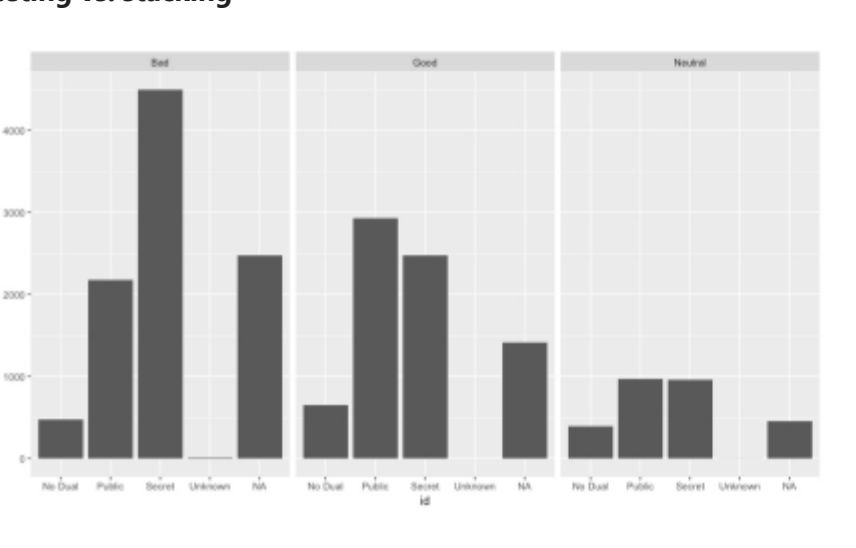
Facetting

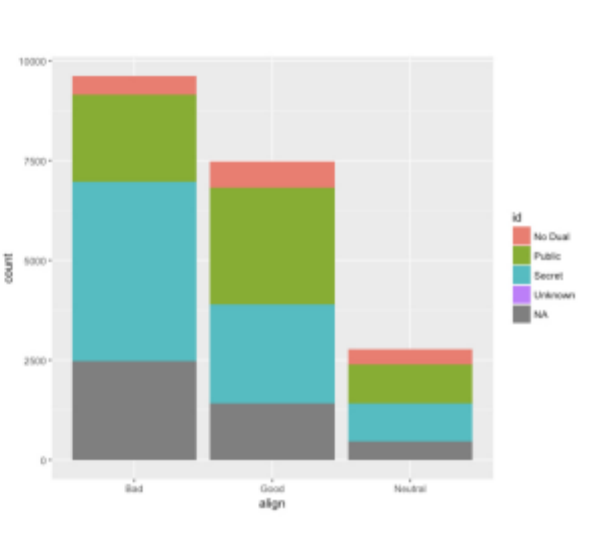
- Break down the data based on the levels of a category

In [52]:
```
# Facetted barchart
ggplot(comics_filtered, aes(x = id)) +
geom_bar() +
facet_wrap(~align)
```



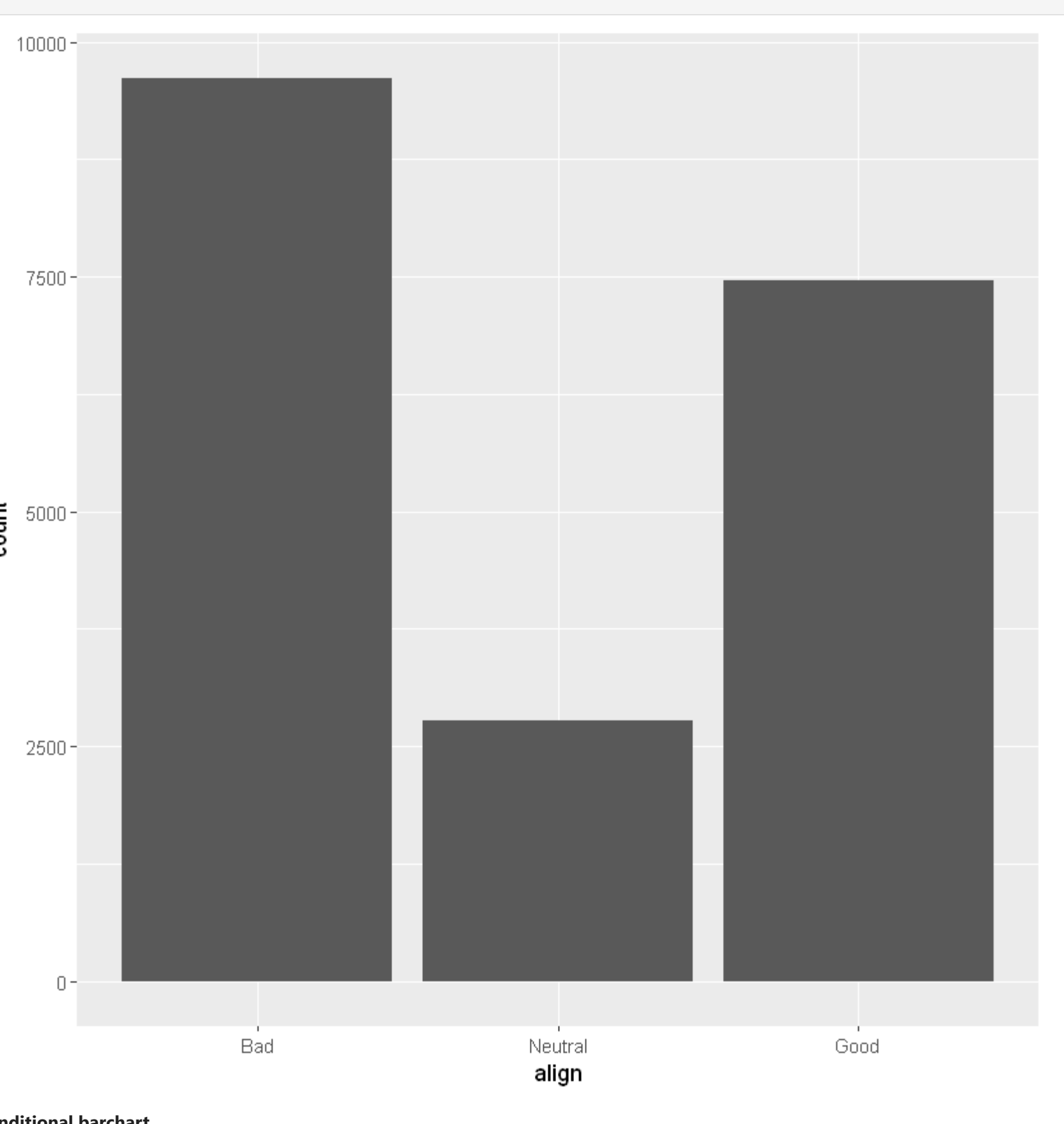
Faceting vs. stacking





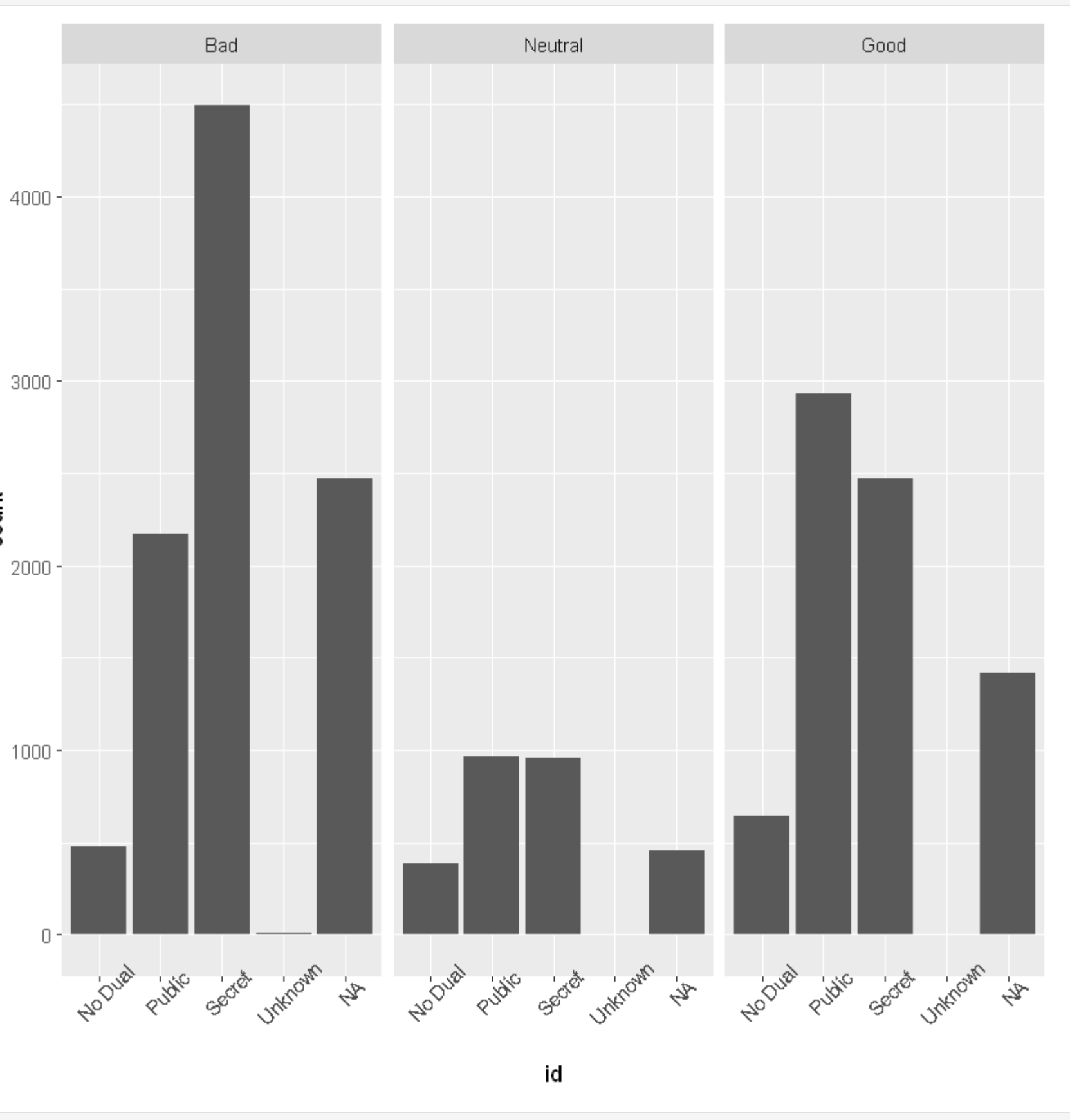
Marginal barchart

In [54]:
```
# Change the order of the levels in align
comics_filtered$align <- factor(comics_filtered$align,
                        levels = c("Bad", "Neutral", "Good"))

# Create plot of align
ggplot(comics_filtered , aes(x = align)) +
  geom_bar()
```



Conditional barchart

In [63]:
```
# Plot of alignment broken down by alignment
ggplot(comics_filtered, aes(x = id)) +
  geom_bar() +
  facet_wrap(~ align) +
    theme(axis.text.x = element_text(angle = 45))
```



In [ ]: