EXPLORATORY DATA ANALYSIS

Exploring categorical data

Comics dataset

```
> comics
# A tibble: 23,272 x 11
                                      name                id   align
                                    <fctr>            <fctr>  <fctr>
1                  Spider-Man (Peter Parker)  Secret Identity    Good
2            Captain America (Steven Rogers)  Public Identity    Good
3  Wolverine (James \\"Logan\\" Howlett)  Public Identity Neutral
4    Iron Man (Anthony \\"Tony\\" Stark)  Public Identity    Good
5                      Thor (Thor Odinson) No Dual Identity    Good
6                Benjamin Grimm (Earth-616)  Public Identity    Good
7                 Reed Richards (Earth-616)  Public Identity    Good
8                Hulk (Robert Bruce Banner)  Public Identity    Good
9                 Scott Summers (Earth-616)  Public Identity Neutral
10               Jonathan Storm (Earth-616)  Public Identity    Good
# ... with 23,262 more rows, and 8 more variables: eye <fctr>,
#   hair <fctr>, gender <fctr>, gsm <fctr>, alive <fctr>,
#   appearances <int>, first_appear <fctr>, publisher <fctr>
```

Working with factors

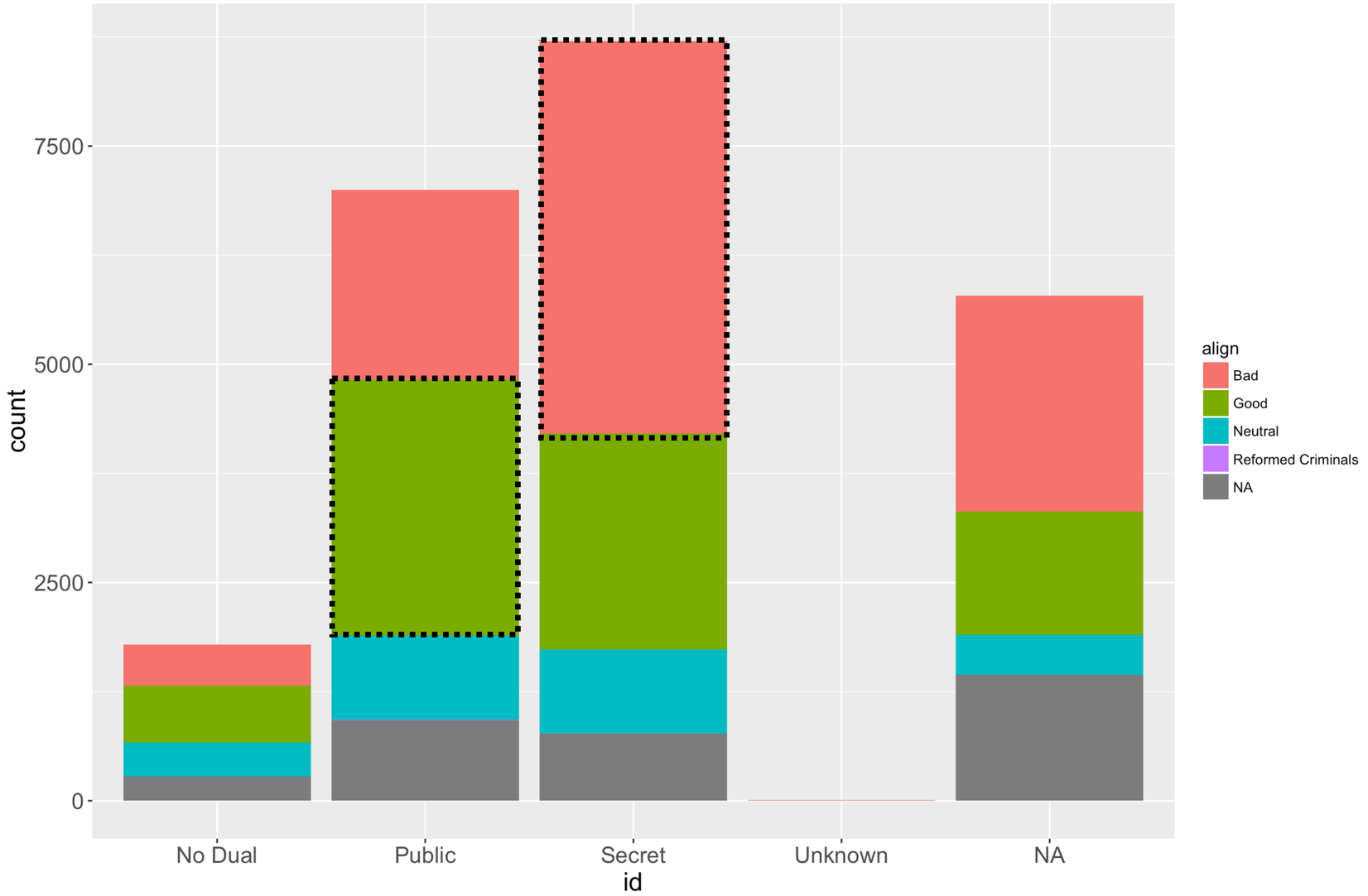
```
> levels(comics$align)
[1] "Bad"                 "Good"                 "Neutral"
[4] "Reformed Criminals"

> levels(comics$id)
[1] "No Dual" "Public"  "Secret"  "Unknown"

> 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
```

Note: NAs ignored by `levels()` function

```
ggplot(data, aes(x = var1, fill = var2)) +
  layer_name()
```

```
ggplot(comics, aes(x = id, fill = align)) +
  geom_bar()
```

Bar chart


7500
5000
2500
0
count
No Dual
Public
Secret
Unknown
NA
id
align
Bad
Good
Neutral
Reformed Criminals
NA

EXPLORATORY DATA ANALYSIS

Let’s practice!

EXPLORATORY DATA ANALYSIS

Counts vs. proportions

From counts to proportions

```
> options(scipen = 999, digits = 3) # Simplify display format
> tab_cnt <- table(comics$id, comics$align)
> tab_cnt
           Bad Good Neutral
  No Dual  474  647     390
  Public  2172 2930     965
  Secret  4493 2475     959
  Unknown    7    0       2

> prop.table(tab_cnt)
                Bad     Good  Neutral
  No Dual 0.030553 0.041704 0.025139
  Public  0.140003 0.188862 0.062202
  Secret  0.289609 0.159533 0.061815
  Unknown 0.000451 0.000000 0.000129

> sum(prop.table(tab_cnt))
[1] 1
```

Conditional proportions

```
> prop.table(tab_cnt, 1)
```

Condition on the rows (i.e. rows sum to 1)

```
             Bad  Good Neutral
  No Dual 0.314 0.428   0.258
  Public  0.358 0.483   0.159
  Secret  0.567 0.312   0.121
  Unknown 0.778 0.000   0.222
```

```
> prop.table(tab_cnt, 2)
```

Condition on the columns (i.e. columns sum to 1)

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

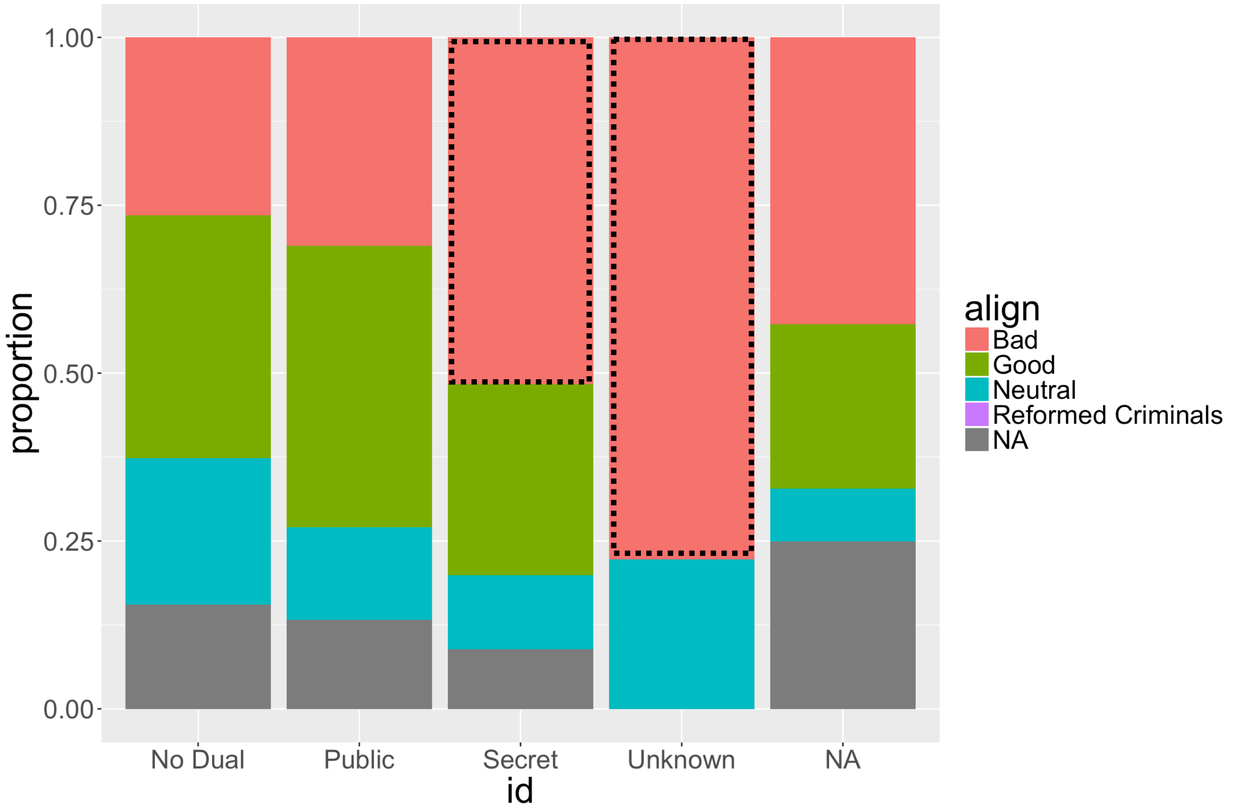
```
ggplot(comics, aes(x = id, fill = align)) +
  geom_bar()
```

```
ggplot(comics, aes(x = id, fill = align)) +
  geom_bar(position = "fill")
```

```
ggplot(comics, aes(x = id, fill = align)) +
  geom_bar(position = "fill") +
  ylab("proportion")
```

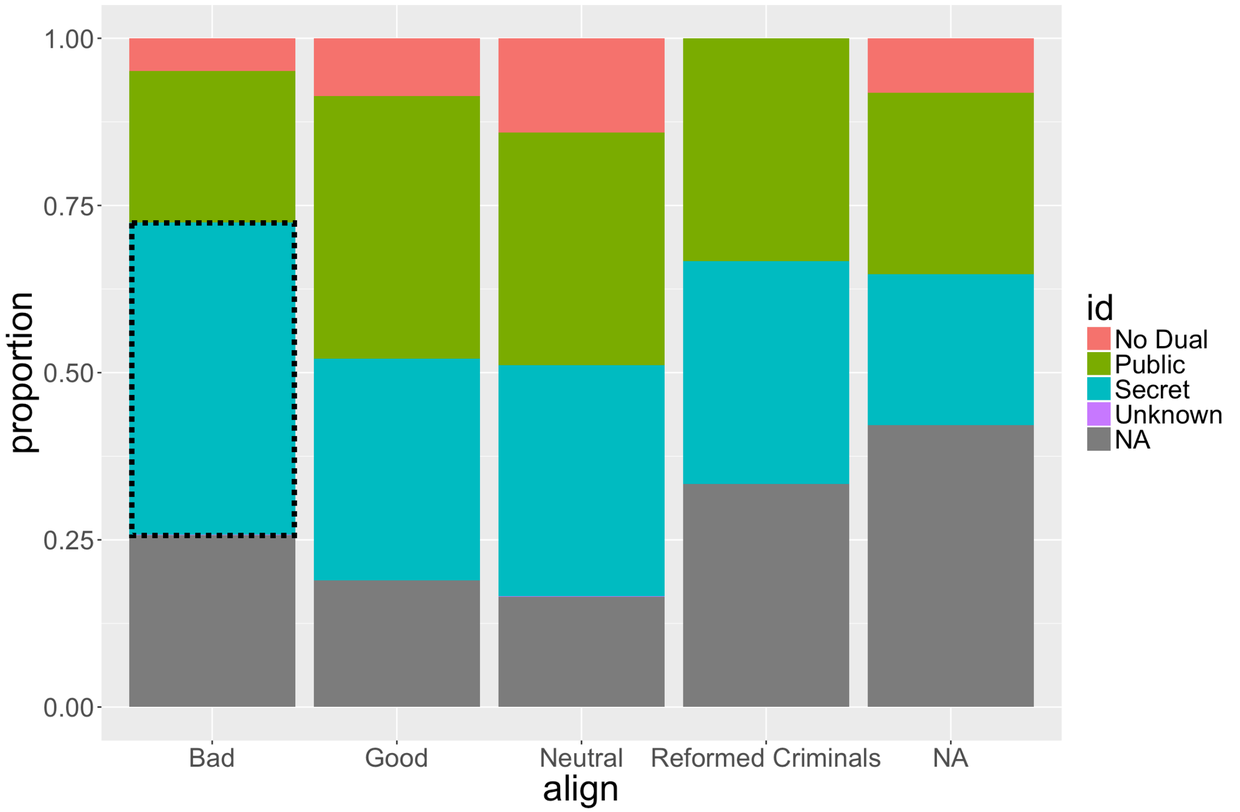
Conditional bar chart

```
> ggplot(comics, aes(x = id, fill = align)) +
    geom_bar(position = "fill") +
    ylab("proportion")
```



Conditional bar chart

```
> ggplot(comics, aes(x = align, fill = id)) +
    geom_bar(position = "fill") +
    ylab("proportion")
```



EXPLORATORY DATA ANALYSIS

Let’s practice!

EXPLORATORY DATA ANALYSIS

Distribution of one variable

Marginal distribution

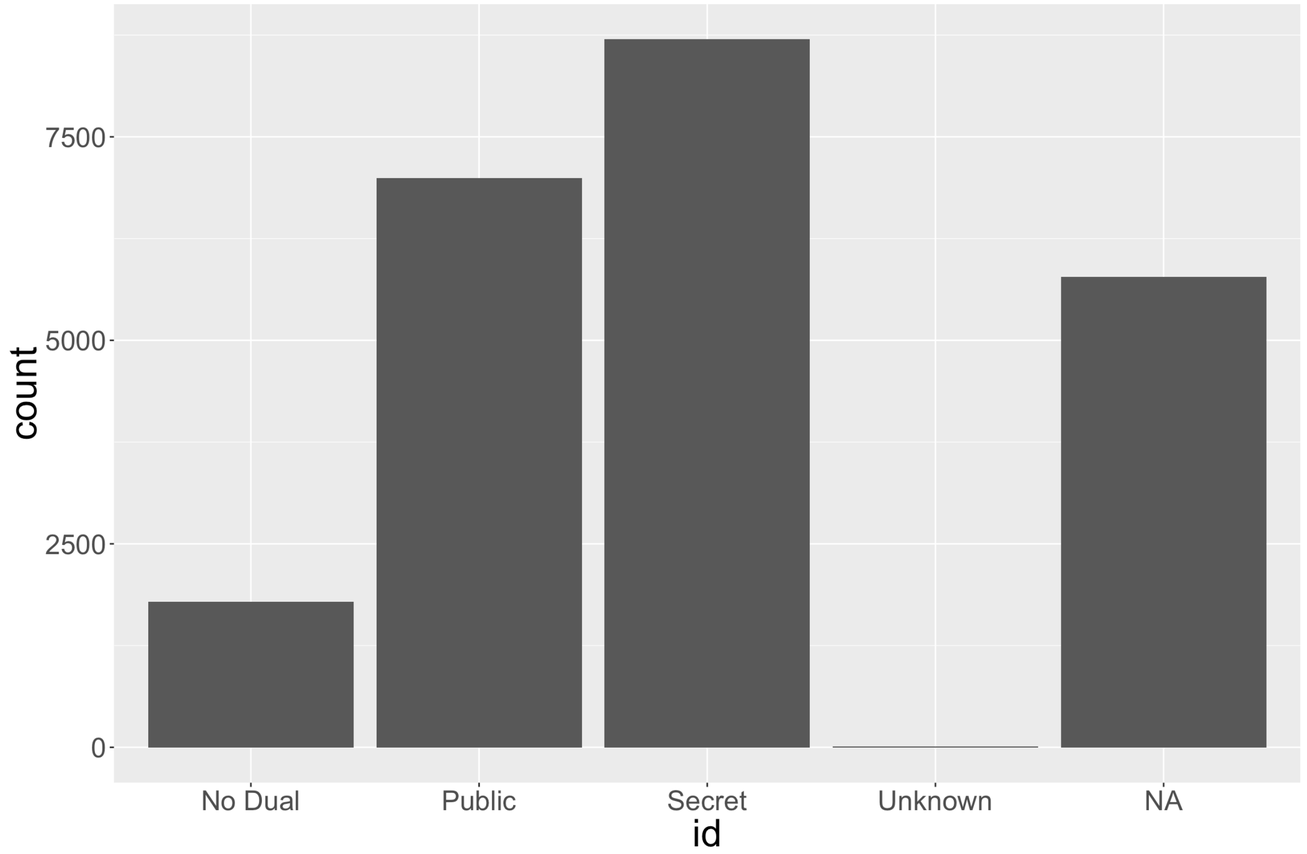
```
> table(comics$id)

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

```
> tab_cnt <- table(comics$id, comics$align)
> tab_cnt
           Bad Good Neutral
No Dual    474  647     390
Public    2172 2930     965
Secret    4493 2475     959
Unknown      7    0       2
```

474 + 647 + 390 = 1511

Simple barchart

```
> ggplot(comics, aes(x = id)) +
    geom_bar()
```

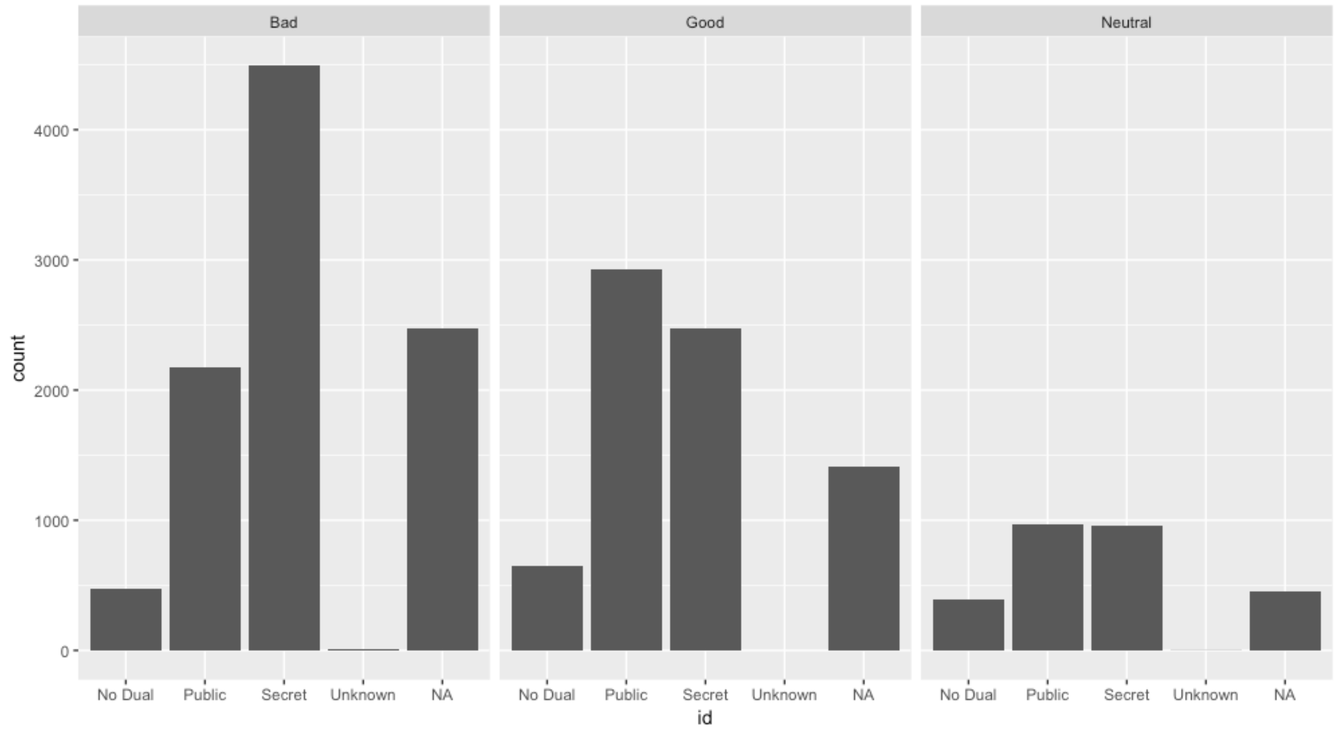


Faceting

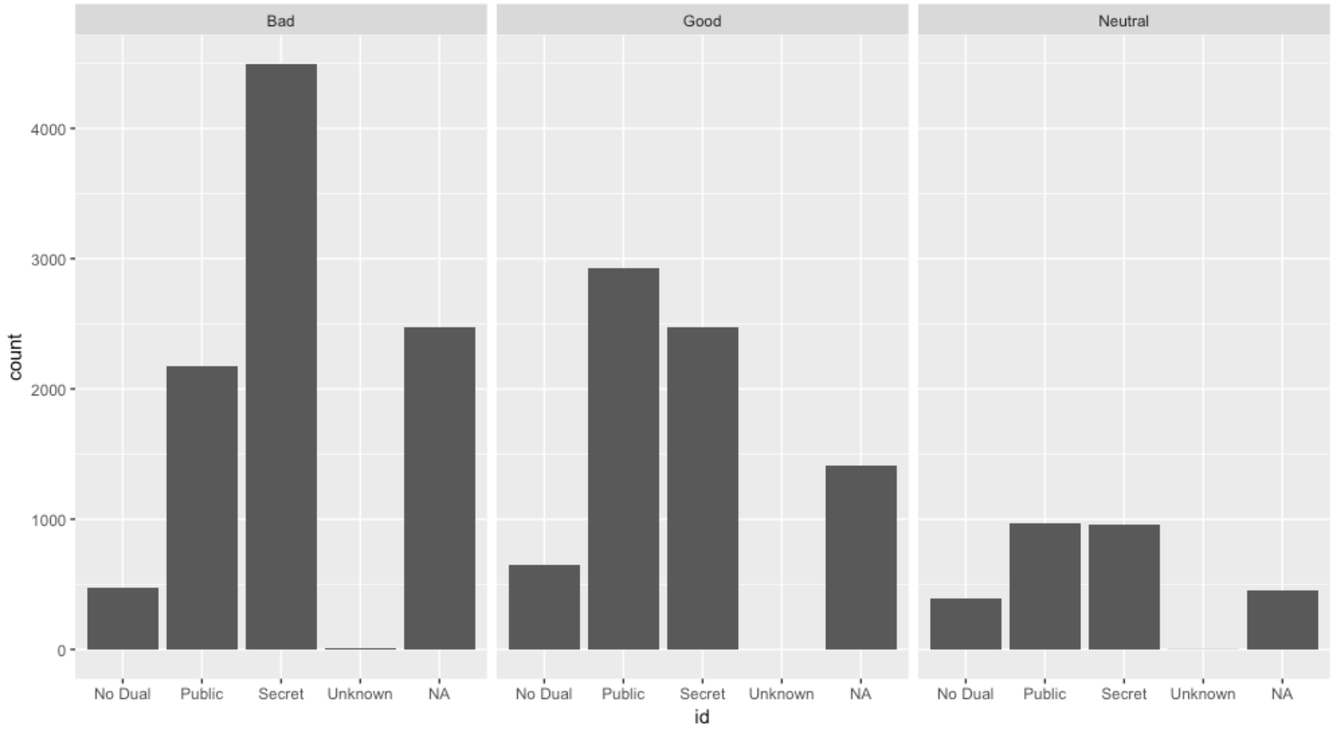
```
> tab_cnt <- table(comics$id, comics$align)
> tab_cnt
           Bad Good Neutral
  No Dual  474  647     390
  Public  2172 2930     965
  Secret  4493 2475     959
  Unknown    7    0       2
```

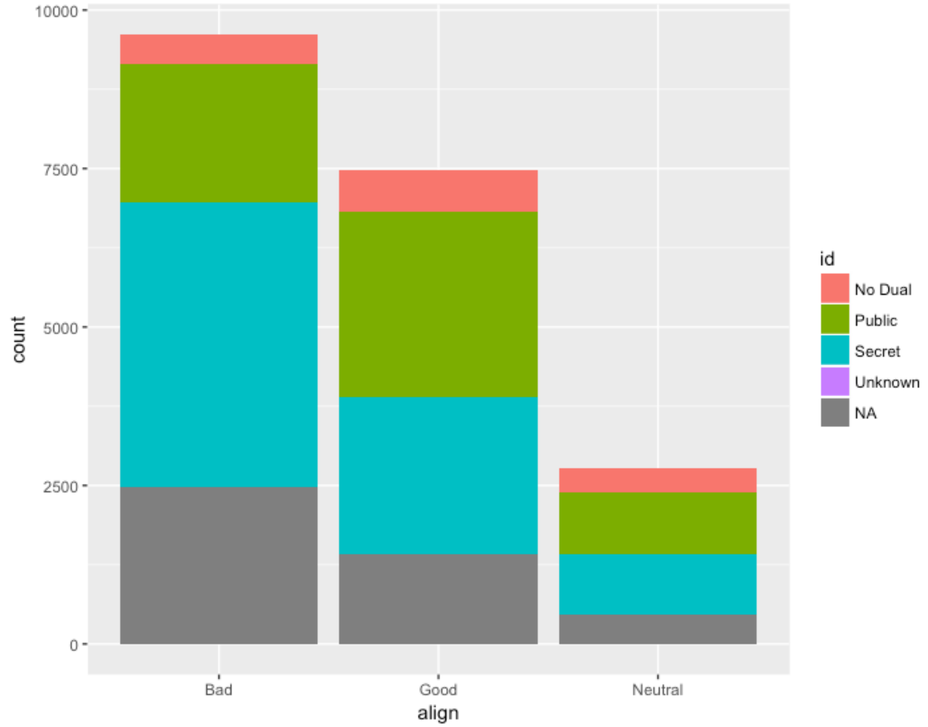
Faceted barcharts

```
> ggplot(comics, aes(x = id)) +
    geom_bar() +
    facet_wrap(~align)
```

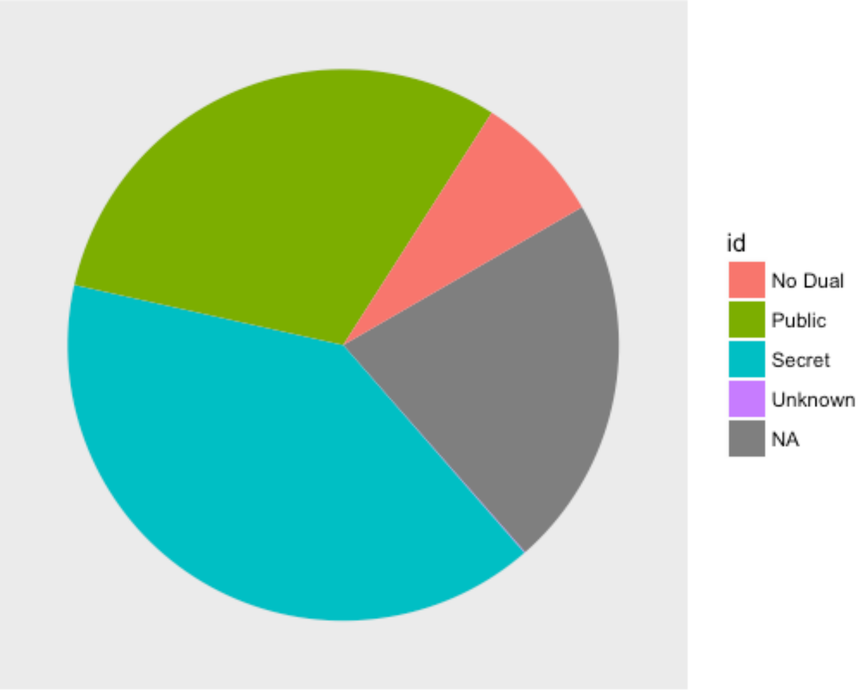


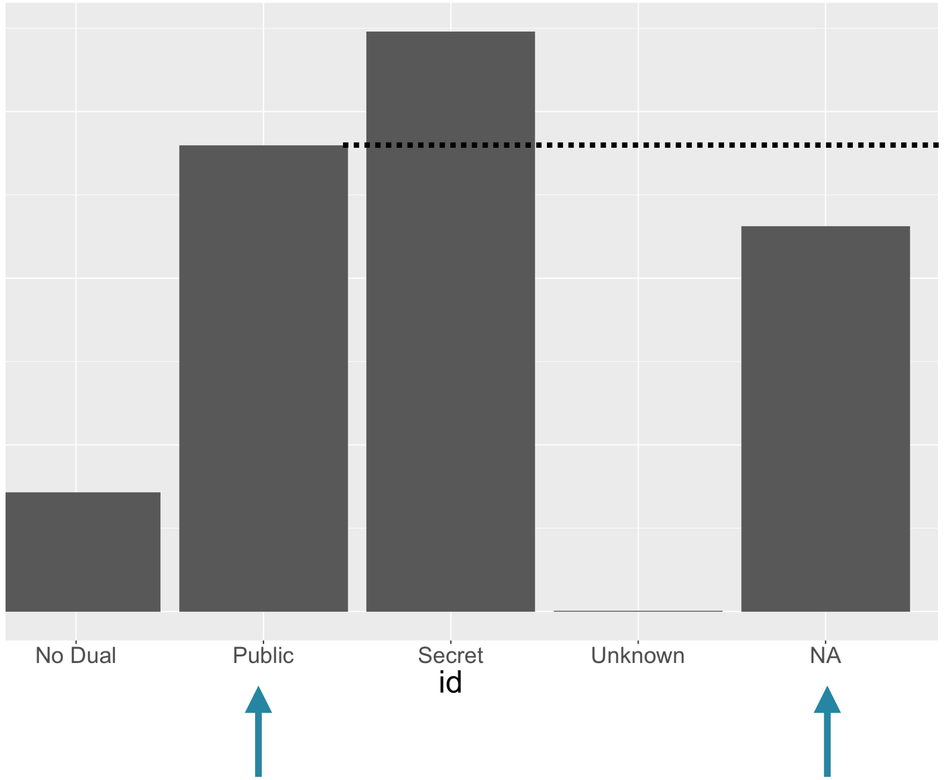
Faceting vs. stacking





Pie chart vs. bar chart





EXPLORATORY DATA ANALYSIS

Let’s practice!