IGN Data: Applying Statistics

by Ellen A. Savoye

Get going by asking the following questions and looking for the answers with some code and plots:

Can you count something interesting?

Can you find some trends (high, low, increase, decrease, anomalies)?

Can you make a bar plot or a histogram?

Can you compare two related quantities?

Can you make a scatterplot?

Can you make a time-series plot?

Having made these plots, what are some insights you get from them? Do you see any correlations? Is there a hypothesis you would like to investigate further? What other questions do they lead you to ask?

After cleaning up my dataset, I moved into applying some exploratory data analysis to unearth some inferences about the data.

```r
head(tbl_df(IGN_data), 5)
```

```
## # A tibble: 5 x 14
##     X.1     X score_phrase title url   platform score genre editors_choice
##   <int> <int> <fct>        <fct> <fct> <fct>    <dbl> <fct> <fct>
## 1     1     0 Amazing      Litt~ /gam~ PlaySta~  9.00 Plat~ Y
## 2     2     1 Amazing      Litt~ /gam~ PlaySta~  9.00 Plat~ Y
## 3     3     2 Great        Spli~ /gam~ iPad      8.50 Puzz~ N
## 4     4     3 Great        NHL ~ /gam~ Xbox 360  8.50 Spor~ N
## 5     5     4 Great        NHL ~ /gam~ PlaySta~  8.50 Spor~ N
## # ... with 5 more variables: release_year <int>, release_month <int>,
## #   release_day <int>, platform_group <fct>, genre_group <fct>
```

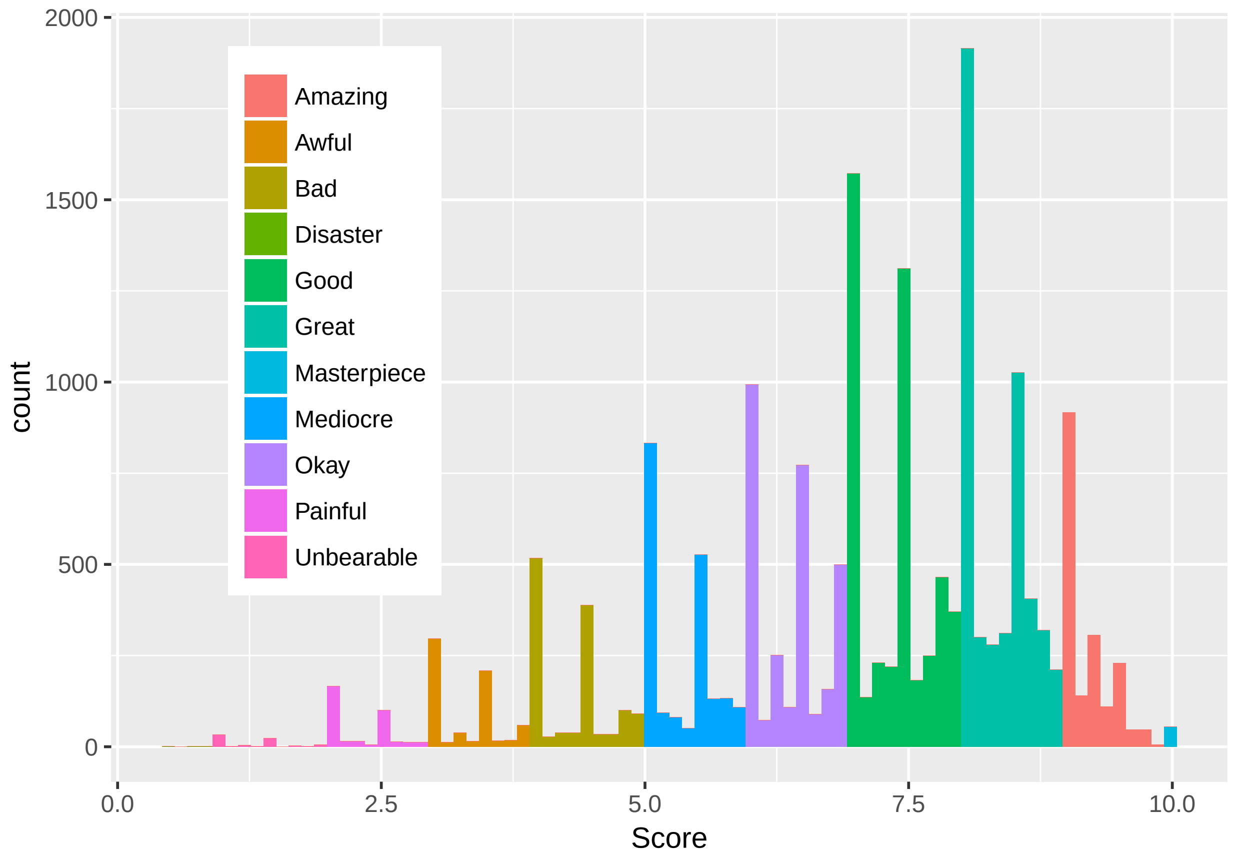
```r
top10 <- head(names((sort(table(IGN_data$genre_group), decreasing = TRUE))), 10)
```

The plots below are based on all 20 years worth of data. I knew the more positive the score phrase, the better the score. However, I wanted to plot see how far down an “Editor’s Choice” game would score. Based on the second graph below, the lowest is approximately an 8. However intersting this fact is, it only goes to show how large the score range is for an “Editor’s Choice” game can be.

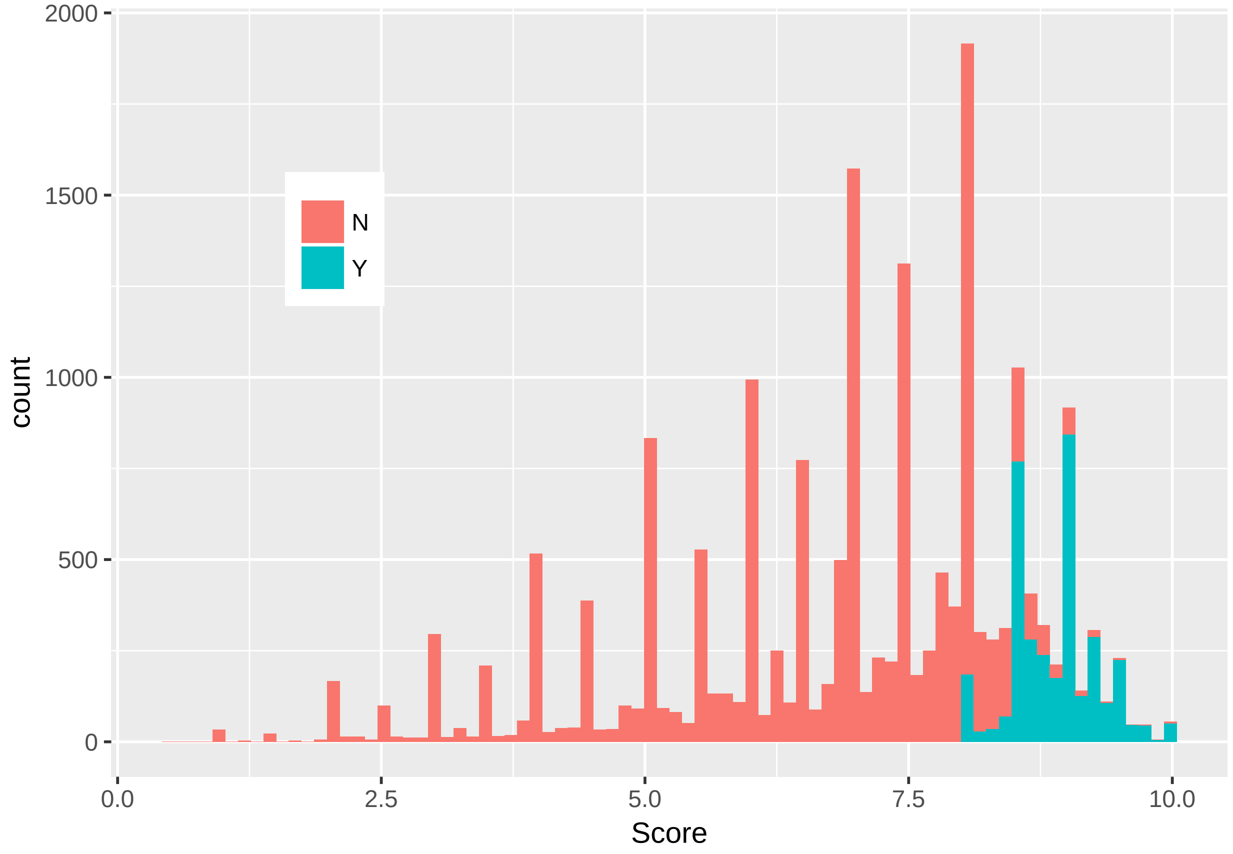
```r
sort(unique(IGN_data$score_phrase))
```

```
##  [1] Amazing     Awful       Bad         Disaster    Good
##  [6] Great       Masterpiece Mediocre    Okay        Painful
## [11] Unbearable
## 11 Levels: Amazing Awful Bad Disaster Good Great Masterpiece ... Unbearable
```

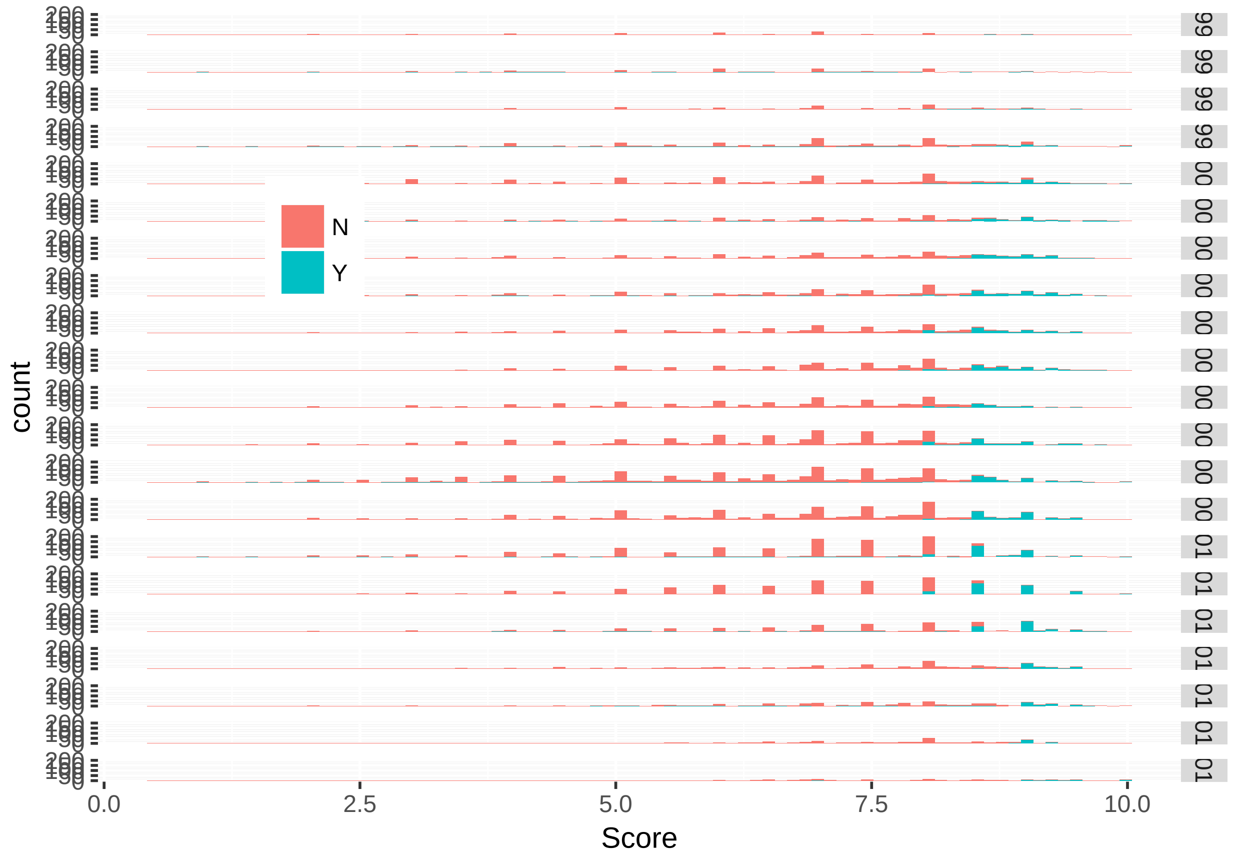
```r
#Score_vs_Phrase
ggplot(data=IGN_data, aes(score)) +
  geom_histogram(aes(fill=factor(score_phrase)),bins=80) +
  xlab('Score') +
  theme(legend.position=c(.2, .6)) +
  theme(legend.title=element_blank())
```



```r
#Score_vs_EdChc
ggplot(data=IGN_data, aes(score)) +
  geom_histogram(aes(fill=factor(editors_choice)),bins=80) +
  xlab('Score') + theme(legend.position=c(.2, .7)) +
  theme(legend.title=element_blank())
```



```
#Score_vs_EdChc_vs_Yr
ggplot(data=IGN_data, aes(score)) +
  geom_histogram(aes(fill=factor(editors_choice)),bins=80) +
  xlab('Score') + theme(legend.position=c(.2, .7)) +
  theme(legend.title=element_blank()) +
  facet_grid(IGN_data$release_year ~ ., IGN_data$release_year > 2007)
```

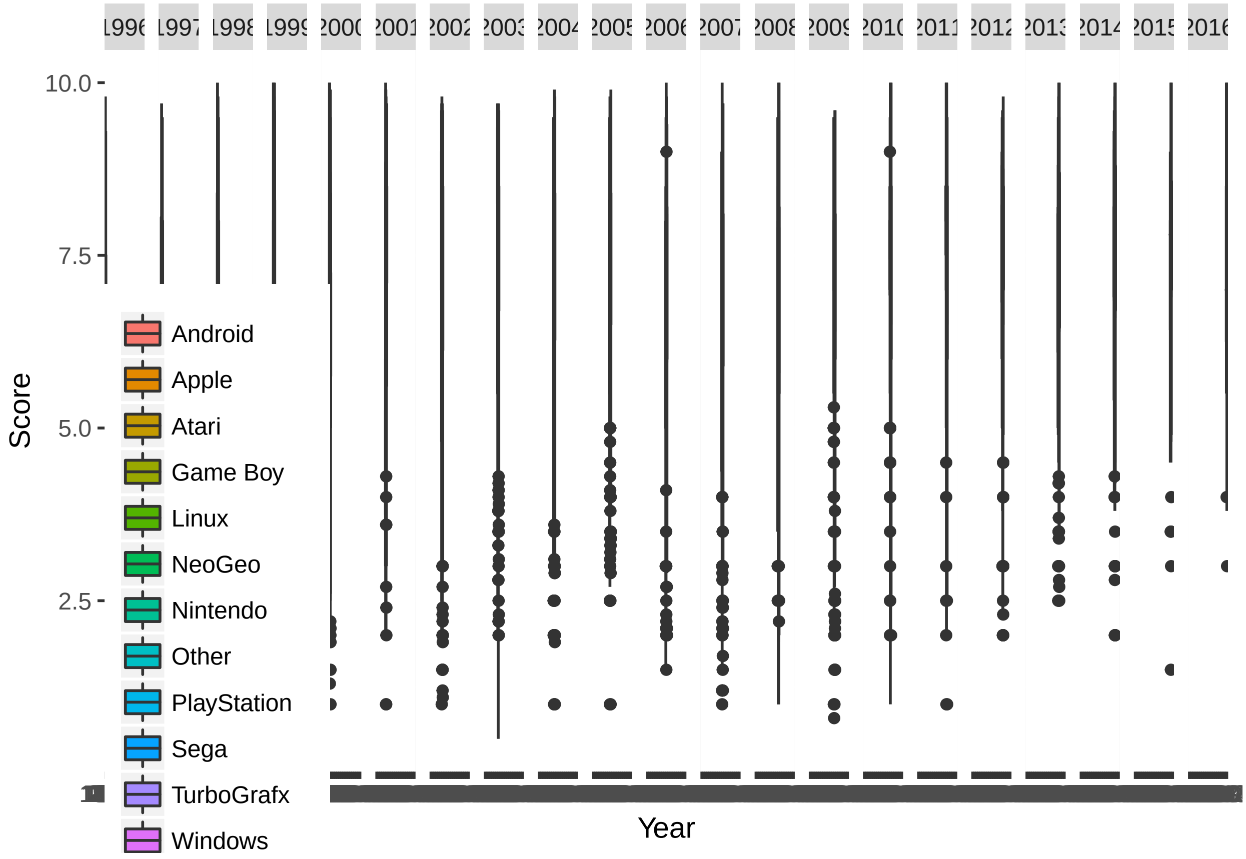


Comparing score versus the cleaned up platform group, I'd like to see if the scores fluctuated for each group with more than 100 records over the last 20 years.

```
IGN_data %>%
  group_by(platform_group) %>%
  summarise(no_rows = length(platform_group)) %>%
  arrange(no_rows)
```

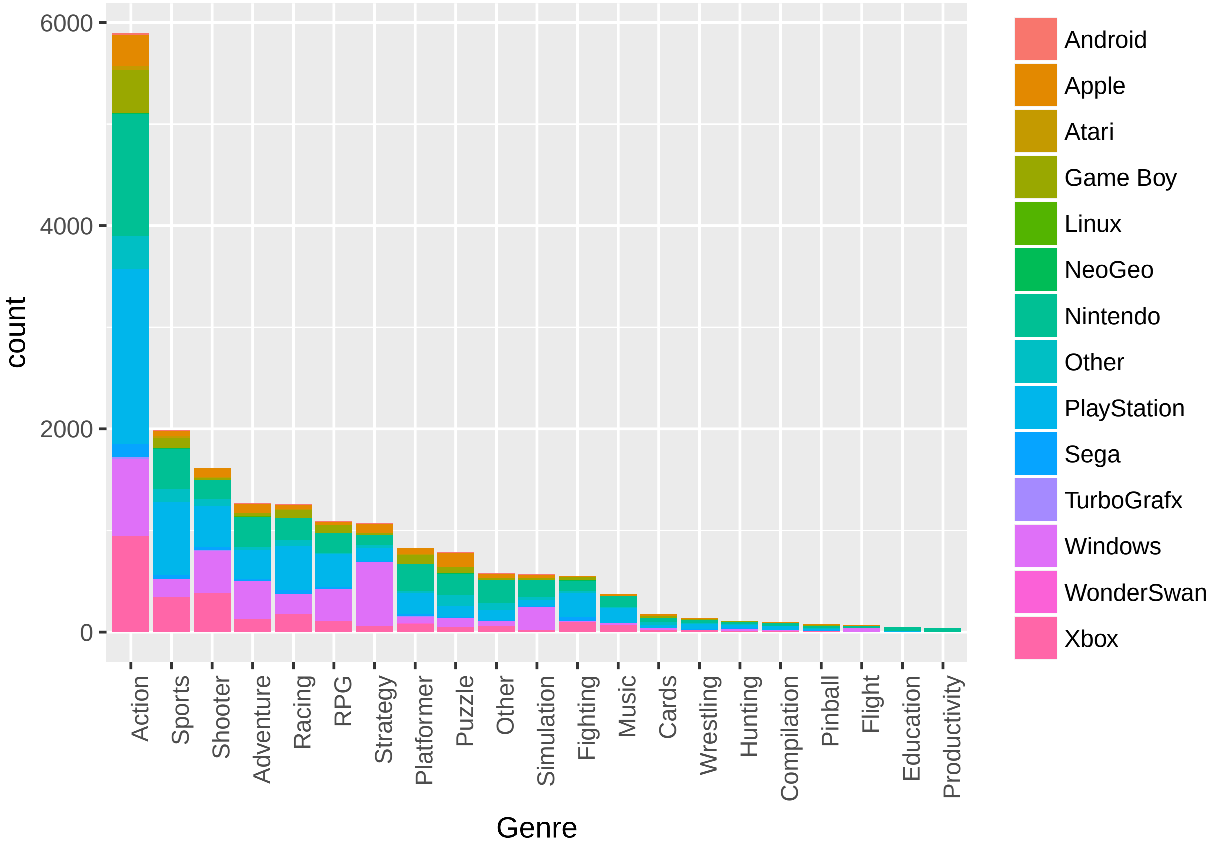
```
## # A tibble: 14 x 2
##    platform_group no_rows
##    <fct>            <int>
##  1 WonderSwan           5
##  2 Linux               11
##  3 Android             40
##  4 NeoGeo              41
##  5 TurboGrafx          43
##  6 Atari               89
##  7 Sega               381
##  8 Other              964
##  9 Game Boy          1001
## 10 Apple             1039
## 11 Xbox              2660
## 12 Windows           3386
## 13 Nintendo          3906
## 14 PlayStation       5059
```

```
ggplot(IGN_data, aes(x = factor(release_year), y = score)) +
  geom_boxplot(aes(fill=(platform_group))) +
  theme(legend.position=c(.1, .2)) +
  xlab('Year') +
  ylab('Score') +
  theme(legend.title=element_blank()) +
  facet_grid(. ~ release_year)
```

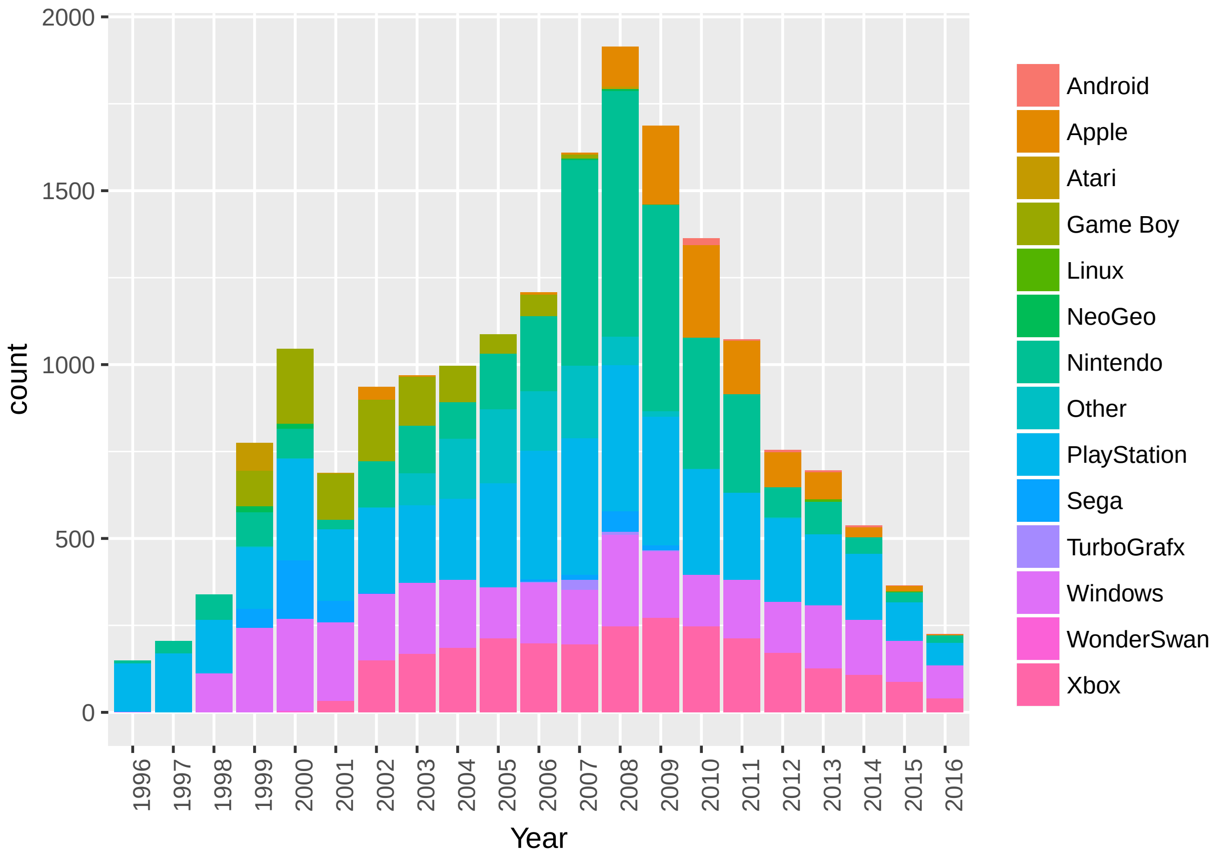


The bar plot below serves two functions. The first is to see the magnitude of each genre. The second is to determine if there is a system that is more focused on releasing certain genres of games. Action, Sports, and Shooter games are the most popular. However, Nintendo has less presence in the Shooter game division. THe second graph below shows Platform versus year. 2008 was a very busy year. However, this can be accounted for as the year Ninentdo Wii came out.

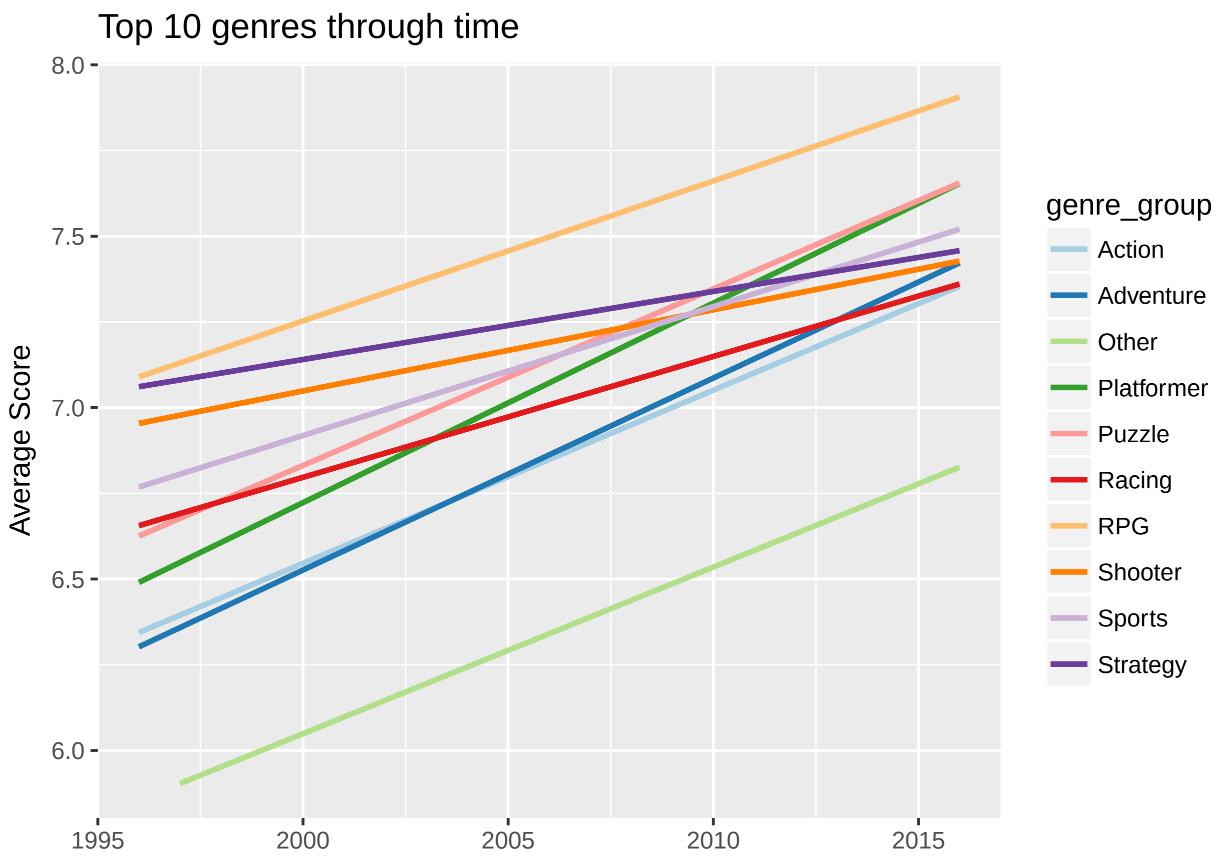
```
#Platform vs Genre
ggplot(IGN_data, aes(x=reorder(genre_group,genre_group,function(x)-length(x)))) +
  geom_bar(aes(fill=platform_group)) +
  theme(axis.text.x = element_text(angle=90, hjust=1)) +
  xlab('Genre') + theme(legend.title=element_blank())
```



```
#Platform vs Year
ggplot(IGN_data, aes(x=factor(release_year))) +
  geom_bar(aes(fill=platform_group)) +
  theme(axis.text.x = element_text(angle=90, hjust=1)) +
  xlab('Year') + theme(legend.title=element_blank())
```



```
IGN_data %>%
  filter(genre_group %in% top10) %>%
  group_by(genre_group, release_year) %>%
  summarize(average_score = mean(score, na.rm = TRUE)) %>%
  ggplot(aes(x = release_year, y = average_score, col = genre_group)) +
    geom_smooth(method = "lm", se = FALSE) +
    theme(axis.title.x = element_blank()) +
    labs(y = "Average Score", title = "Top 10 genres through time", x ="") +
    scale_color_brewer(palette = "Paired")
```



Count something interesting: # of editors choice games, score per genre,

Trends: day of week release, time of the year,

Bar plot or histogram:

scatterplot:

time-series plot:

Having made these plots, what are some insights you get from them? Do you see any correlations? Is there a hypothesis you would like to investigate further? What other questions do they lead you to ask?