

Final Project - Draft

Suthi de Silva

1/22/2024

1: Cleaning the data!

Some of the user reviews are “tbd”, meaning “to be decided”. To keep comparisons (particularly between user and meta scores) equal, I excluded them from the data set.

```
allgamescopy = allgamescopy[allgamescopy$user_review != "tbd", ]
```

user_review’s type is chr, meaning it will not work as a quantitative variable. To navigate this, I made a new column called user_score, and assigned the values there as a numeric type.

```
allgamescopy$user_score <- 0

for (i in 1:nrow(allgamescopy)){
  allgamescopy[i, "user_score"] <- as.numeric(allgamescopy[i, "user_review"])
}
```

For consistency, I multiplied the user scores by 10 so that they will be out of 100, like the meta scores.

```
for (i in 1:nrow(allgamescopy)) {
  allgamescopy[i, "user_score"] <- (allgamescopy[i, "user_score"] * 10)
}
```

I also made a column for the user and meta score combined, called total_score. This should be measured out of 200.

```
allgamescopy$total_score <- 0

for (i in 1:nrow(allgamescopy)) {
  allgamescopy[i, "total_score"] <- (allgamescopy[i, "meta_score"] +
                                     allgamescopy[i, "user_score"])
}
```

The platform column has whitespace at the start of each variable. I removed it.

```
for (i in 1:nrow(allgamescopy)) {
  allgamescopy[i, "platform"] <- str_sub(allgamescopy[i, "platform"], 2)
}
```

The release_date column gives the entire date. For categorisation purposes, I made a column for just the release year specifically.

```
allgamescopy$year <- ""

for (i in 1:nrow(allgamescopy)) {
  allgamescopy[i, "year"] <- str_sub(allgamescopy[i, "release_date"], -4, -1)
}
```

I made a new column to identify each game by the game generation in which it was released.

5th Generation: 1993-1997 | 6th Generation: 1998-2004 | 7th Generation: 2005-2011 | 8th Generation: 2012-2019 | 9th Generation: 2020 onward (current generation)

```
allgamescopy$generation <- " "  
  
for (i in 1:nrow(allgamescopy)) {  
  if (allgamescopy[i, "year"] <= 1997) {  
    allgamescopy[i, "generation"] <- "5th Generation"  
  }  
  else if (allgamescopy[i, "year"] <= 2004) {  
    allgamescopy[i, "generation"] <- "6th Generation"  
  }  
  else if (allgamescopy[i, "year"] <= 2011) {  
    allgamescopy[i, "generation"] <- "7th Generation"  
  }  
  else if (allgamescopy[i, "year"] <= 2019) {  
    allgamescopy[i, "generation"] <- "8th Generation"  
  }  
  else if (allgamescopy[i, "year"] >= 2020) {  
    allgamescopy[i, "generation"] <- "9th Generation"  
  }  
}
```

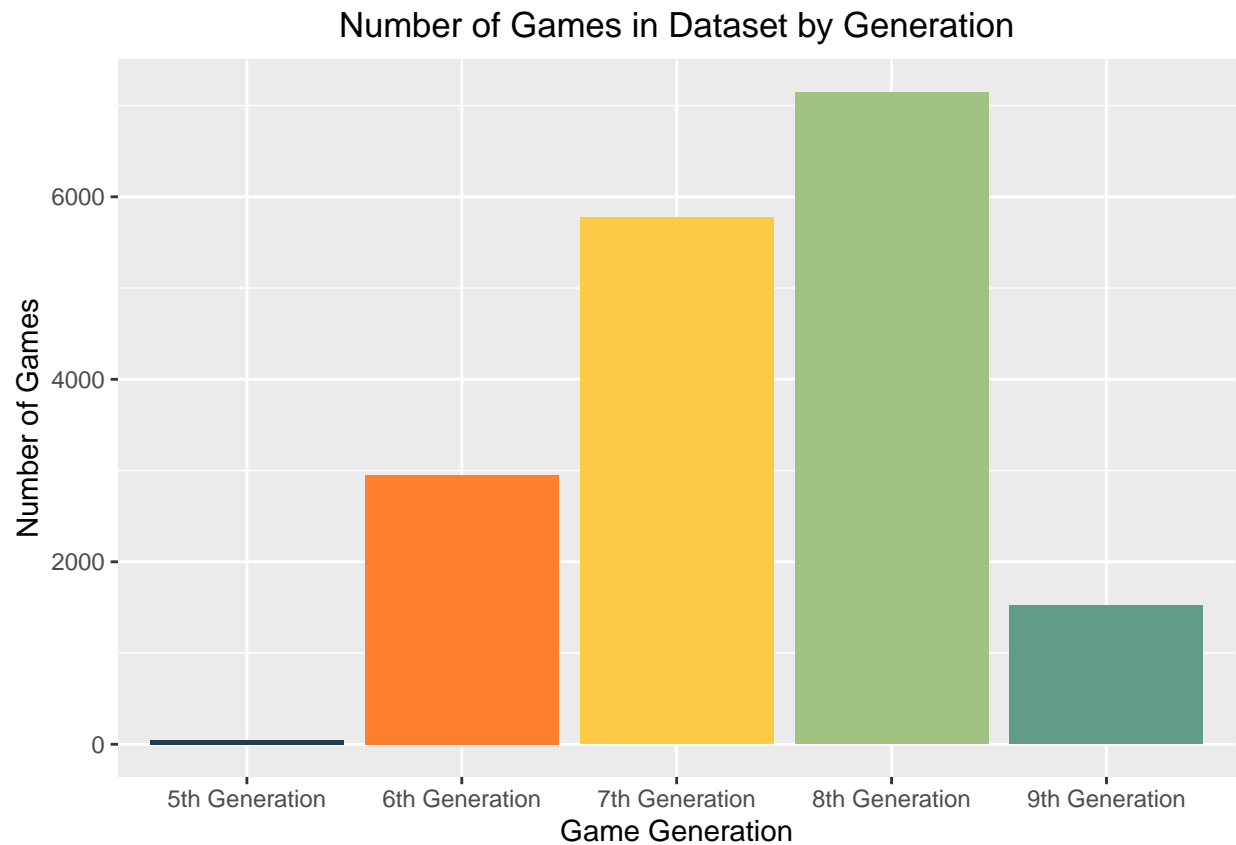
I categorised what I considered to be relevant columns, for grouping purposes.

```
allgamescopy$platform <- as.factor(allgamescopy$platform)  
allgamescopy$year <- as.factor(allgamescopy$year)  
allgamescopy$generation <- as.factor(allgamescopy$generation)
```

And finally, I made a visual of the number of games per generation and platform in the data set, just so we could easily see what we're working with.

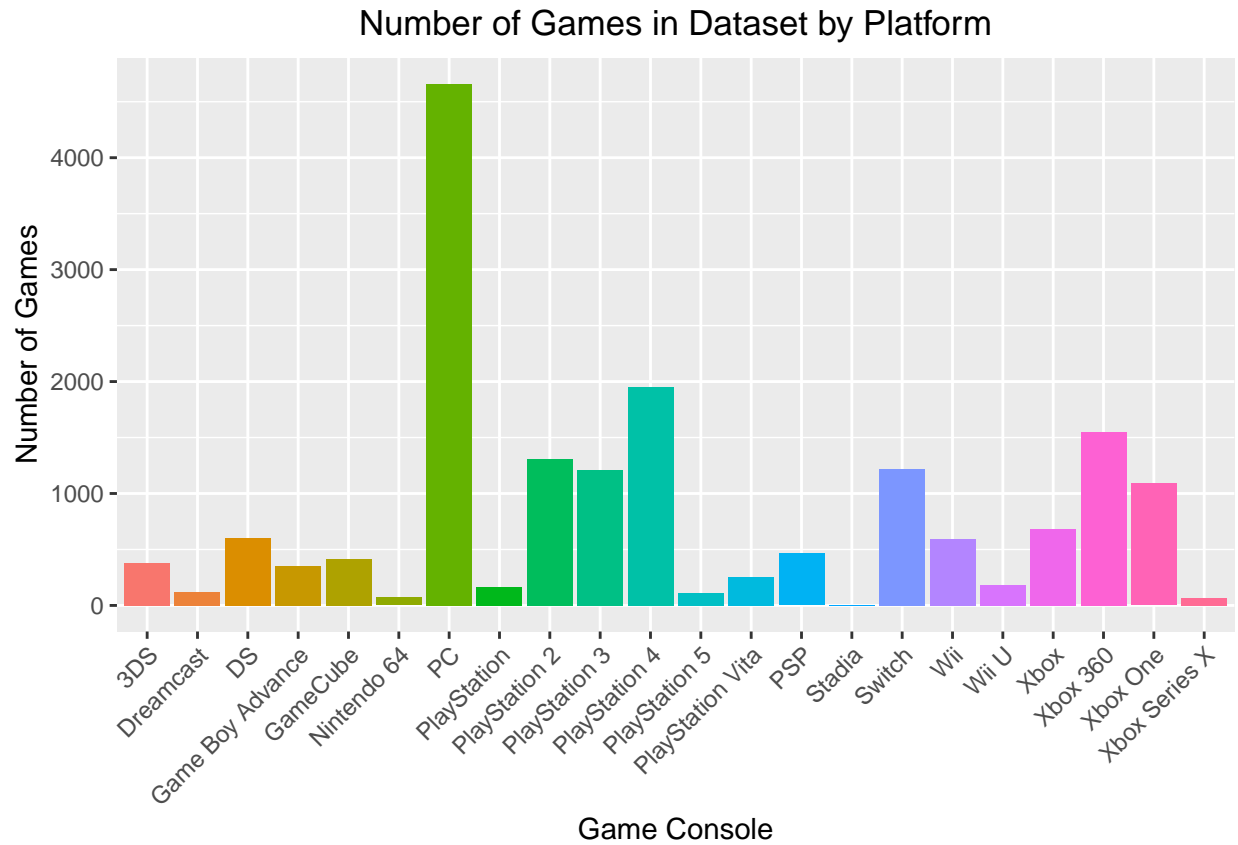
Total Number of Games by Generation

```
ggplot(allgamescopy) + geom_bar(mapping = aes(x=generation, fill=generation)) +  
  labs(title="Number of Games in Dataset by Generation",  
       x="Game Generation", y="Number of Games") +  
  theme(plot.title=element_text(hjust = 0.5), legend.position="none") +  
  scale_fill_manual(values=c("#233D4D", "#FE7F2D", "#FCCA46", "#A1C181", "#619B8A"))
```



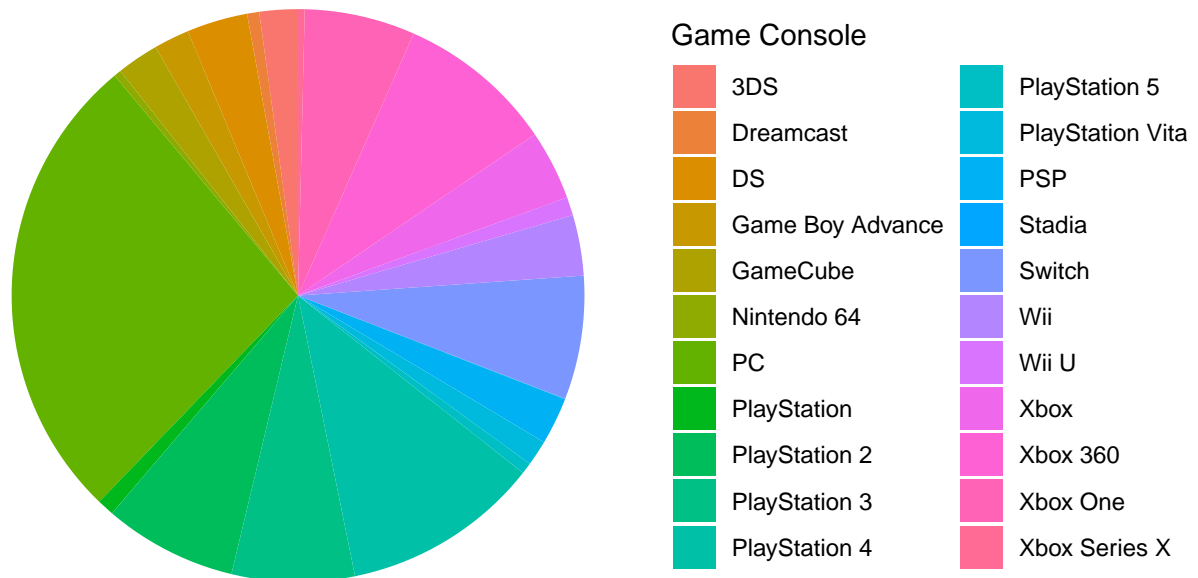
Total Number of Games by Platform

```
ggplot(allgamescopy) + geom_bar(mapping = aes(x=platform, fill=platform)) +  
  labs(title="Number of Games in Dataset by Platform",  
        x="Game Console", y="Number of Games") +  
  theme(plot.title=element_text(hjust = 0.5), legend.position="none") +  
  theme(axis.text.x = element_text(angle=45, vjust=1, hjust=1))
```



```
ggplot(allgamescopy, aes(x=factor(1), fill=platform)) +
  geom_bar(width = 1) +
  coord_polar("y") +
  theme_void() +
  labs(title="Number of Games in Dataset by Platform") +
  theme(plot.title=element_text(hjust = 0.5)) +
  guides(fill = guide_legend(title = "Game Console"))
```

Number of Games in Dataset by Platform



Sanity Check

```
summary(allgamescopy)
```

```
##      index      name      platform      release_date
##  Min.   : 0      Length:17435      PC           :4660      Length:17435
##  1st Qu.: 4404    Class :character      PlayStation 4:1950    Class :character
##  Median : 8958    Mode  :character      Xbox 360       :1547    Mode  :character
##  Mean   : 9120
##  3rd Qu.:13768
##  Max.   :18799
##
##      summary      meta_score      user_review      user_score
##  Length:17435      Min.   :20.0      Length:17435      Min.   : 2.00
##  Class :character  1st Qu.:64.0      Class :character  1st Qu.:63.00
##  Mode  :character  Median :73.0      Mode  :character  Median :73.00
##                      Mean   :71.2      Mean   :69.91
##                      3rd Qu.:80.0      3rd Qu.:79.00
##                      Max.   :99.0      Max.   :97.00
##
##      total_score      year      generation
##  Min.   : 35.0      2018   : 1091      5th Generation: 49
##  1st Qu.:129.0      2017   :  994      6th Generation:2950
##  Median :145.0      2019   :  974      7th Generation:5772
##  Mean   :141.1      2020   :  957      8th Generation:7145
##  3rd Qu.:157.0      2016   :  944      9th Generation:1519
```

```
## Max.      :190.0   2009   : 866
##                               (Other):11609
```

Looking at the data post-wrangling, I can see that the platform, year, and generation columns are now divided into categories per unique variable. This will allow me to use them as quantitative variables in my graphs. The index, name, release date, and summary column are likely irrelevant now and could probably be removed to make the data set cleaner. The user review column could probably also be removed too, since there is now the user score column, deeming the former redundant. Removing the white space from the platform variables was the trickiest thing to catch as it was hard to notice, but removing it at least allows me to use these variables properly. "Switch" and "Switch" are two entirely different variables, and trying to use the latter when the data set has the former led to errors and confusion.

Overall, from the data set, I can see an index number for each variable, the name of each game included, the platform (or console) that each game is on, the date (and specifically the year!) that the game was released, the Metacritic and user scores for the game, a combined score that represents both, and the game generation to which the game belongs.

2: Preparing data for graphing!

I found the average user/meta/total score per generation and grouped them. User and meta scores are out of 100, total score is out of 200.

```
# meta score averages
ninthgenmetaavg <- round(mean(allgamescopy[allgamescopy$generation == '9th Generation',
                                          'meta_score']), digits=2)
eighthgenmetaavg <- round(mean(allgamescopy[allgamescopy$generation == '8th Generation',
                                          'meta_score']), digits=2)
seventhgenmetaavg <- round(mean(allgamescopy[allgamescopy$generation == '7th Generation',
                                          'meta_score']), digits=2)
sixthgenmetaavg <- round(mean(allgamescopy[allgamescopy$generation == '6th Generation',
                                          'meta_score']), digits=2)
fifthgenmetaavg <- round(mean(allgamescopy[allgamescopy$generation == '5th Generation',
                                          'meta_score']), digits=2)

#user score averages
ninthgenuseravg <- round(mean(allgamescopy[allgamescopy$generation == '9th Generation',
                                          'user_score']), digits=2)
eighthgenuseravg <- round(mean(allgamescopy[allgamescopy$generation == '8th Generation',
                                          'user_score']), digits=2)
seventhgenuseravg <- round(mean(allgamescopy[allgamescopy$generation == '7th Generation',
                                          'user_score']), digits=2)
sixthgenuseravg <- round(mean(allgamescopy[allgamescopy$generation == '6th Generation',
                                          'user_score']), digits=2)
fifthgenuseravg <- round(mean(allgamescopy[allgamescopy$generation == '5th Generation',
                                          'user_score']), digits=2)

# combined score averages
ninthgentotalavg <- round(mean(allgamescopy[allgamescopy$generation == '9th Generation',
                                          'total_score']), digits=2)
eighthgentotalavg <- round(mean(allgamescopy[allgamescopy$generation == '8th Generation',
                                          'total_score']), digits=2)
seventhgentotalavg <- round(mean(allgamescopy[allgamescopy$generation == '7th Generation',
                                          'total_score']), digits=2)
sixthgentotalavg <- round(mean(allgamescopy[allgamescopy$generation == '6th Generation',
                                          'total_score']), digits=2)
```

```
fifthgentotalavg <- round(mean(allgamescopy[allgamescopy$generation == '5th Generation',
                                         'total_score']), digits=2)

avgcoregen = data.frame(generation=c('5th Generation', '6th Generation',
                                       '7th Generation', '8th Generation',
                                       '9th Generation'),
                        avg_user_score=c(fifthgenuseravg, sixthgenuseravg,
                                         seventhgenuseravg, eighthgenuseravg,
                                         ninthgenuseravg),
                        avg_meta_score=c(fifthgenmetaavg, sixthgenmetaavg,
                                         seventhgenmetaavg, eighthgenmetaavg,
                                         ninthgenmetaavg),
                        avg_total_score=c(fifthgentotalavg, sixthgentotalavg,
                                         seventhgentotalavg, eighthgentotalavg,
                                         ninthgentotalavg))
```

```
summary(allgamescopy$platform)
```

```
##           3DS           Dreamcast           DS Game Boy Advance
##           378           119           599           349
##      GameCube      Nintendo 64           PC           PlayStation
##           413           71           4660           166
##      PlayStation 2      PlayStation 3      PlayStation 4      PlayStation 5
##           1311           1208           1950           110
## PlayStation Vita           PSP           Stadia           Switch
##           251           464           5           1216
##           Wii           Wii U           Xbox           Xbox 360
##           597           181           686           1547
##      Xbox One      Xbox Series X
##           1089           65
```

```
# meta score averages
threedsmetaavg <- round(mean(allgamescopy[allgamescopy$platform == '3DS',
                                         'meta_score']), digits=2)
nin64metaavg <- round(mean(allgamescopy[allgamescopy$platform == 'Nintendo 64',
                                         'meta_score']), digits=2)
ps4metaavg <- round(mean(allgamescopy[allgamescopy$platform == 'PlayStation 4',
                                         'meta_score']), digits=2)
switchmetaavg <- round(mean(allgamescopy[allgamescopy$platform == 'Switch',
                                         'meta_score']), digits=2)
xbox1metaavg <- round(mean(allgamescopy[allgamescopy$platform == 'Xbox One',
                                         'meta_score']), digits=2)
dreamcastmetaavg <- round(mean(allgamescopy[allgamescopy$platform == 'Dreamcast',
                                         'meta_score']), digits=2)
pcmetaavg <- round(mean(allgamescopy[allgamescopy$platform == 'PC',
                                         'meta_score']), digits=2)
ps5metaavg <- round(mean(allgamescopy[allgamescopy$platform == 'PlayStation 5',
                                         'meta_score']), digits=2)
wiimetaavg <- round(mean(allgamescopy[allgamescopy$platform == 'Wii',
                                         'meta_score']), digits=2)
xboxsxmetaavg <- round(mean(allgamescopy[allgamescopy$platform == 'Xbox Series X',
                                         'meta_score']), digits=2)
dsmetaavg <- round(mean(allgamescopy[allgamescopy$platform == 'DS',
                                         'meta_score']), digits=2)
```

```

psmetaavg <- round(mean(allgamescopy[allgamescopy$platform == 'PlayStation',
                              'meta_score']), digits=2)
psvitametaavg <- round(mean(allgamescopy[allgamescopy$platform == 'PlayStation Vita',
                              'meta_score']), digits=2)
wiiumetaavg <- round(mean(allgamescopy[allgamescopy$platform == 'Wii U',
                              'meta_score']), digits=2)
gbametaavg <- round(mean(allgamescopy[allgamescopy$platform == 'Game Boy Advance',
                              'meta_score']), digits=2)
ps2metaavg <- round(mean(allgamescopy[allgamescopy$platform == 'PlayStation 2',
                              'meta_score']), digits=2)
pspmetaavg <- round(mean(allgamescopy[allgamescopy$platform == 'PSP',
                              'meta_score']), digits=2)
xboxmetaavg <- round(mean(allgamescopy[allgamescopy$platform == 'Xbox',
                              'meta_score']), digits=2)
gamecubemetaavg <- round(mean(allgamescopy[allgamescopy$platform == 'GameCube',
                              'meta_score']), digits=2)
ps3metaavg <- round(mean(allgamescopy[allgamescopy$platform == 'PlayStation 3',
                              'meta_score']), digits=2)
stadiametaavg <- round(mean(allgamescopy[allgamescopy$platform == 'Stadia',
                              'meta_score']), digits=2)
xbox360metaavg <- round(mean(allgamescopy[allgamescopy$platform == 'Xbox 360',
                              'meta_score']), digits=2)

#user score averages
threedsuseravg <- round(mean(allgamescopy[allgamescopy$platform == '3DS',
                              'user_score']), digits=2)
nin64useravg <- round(mean(allgamescopy[allgamescopy$platform == 'Nintendo 64',
                              'user_score']), digits=2)
ps4useravg <- round(mean(allgamescopy[allgamescopy$platform == 'PlayStation 4',
                              'user_score']), digits=2)
switchuseravg <- round(mean(allgamescopy[allgamescopy$platform == 'Switch',
                              'user_score']), digits=2)
xbox1useravg <- round(mean(allgamescopy[allgamescopy$platform == 'Xbox One',
                              'user_score']), digits=2)
dreamcastuseravg <- round(mean(allgamescopy[allgamescopy$platform == 'Dreamcast',
                              'user_score']), digits=2)
pcuseravg <- round(mean(allgamescopy[allgamescopy$platform == 'PC',
                              'user_score']), digits=2)
ps5useravg <- round(mean(allgamescopy[allgamescopy$platform == 'PlayStation 5',
                              'user_score']), digits=2)
wiiuseravg <- round(mean(allgamescopy[allgamescopy$platform == 'Wii',
                              'user_score']), digits=2)
xboxsxuseravg <- round(mean(allgamescopy[allgamescopy$platform == 'Xbox Series X',
                              'user_score']), digits=2)
dsuseravg <- round(mean(allgamescopy[allgamescopy$platform == 'DS',
                              'user_score']), digits=2)
psuseravg <- round(mean(allgamescopy[allgamescopy$platform == 'PlayStation',
                              'user_score']), digits=2)
psvitauseravg <- round(mean(allgamescopy[allgamescopy$platform == 'PlayStation Vita',
                              'user_score']), digits=2)
wiiuuseravg <- round(mean(allgamescopy[allgamescopy$platform == 'Wii U',
                              'user_score']), digits=2)
gbauseravg <- round(mean(allgamescopy[allgamescopy$platform == 'Game Boy Advance',

```



```

        'user_score']], digits=2)
ps2useravg <- round(mean(allgamescopy[allgamescopy$platform == 'PlayStation 2',
        'user_score']], digits=2)
pspuseravg <- round(mean(allgamescopy[allgamescopy$platform == 'PSP',
        'user_score']], digits=2)
xboxuseravg <- round(mean(allgamescopy[allgamescopy$platform == 'Xbox',
        'user_score']], digits=2)
gamecubeuseravg <- round(mean(allgamescopy[allgamescopy$platform == 'GameCube',
        'user_score']], digits=2)
ps3useravg <- round(mean(allgamescopy[allgamescopy$platform == 'PlayStation 3',
        'user_score']], digits=2)
stadiauseravg <- round(mean(allgamescopy[allgamescopy$platform == 'Stadia',
        'user_score']], digits=2)
xbox360useravg <- round(mean(allgamescopy[allgamescopy$platform == 'Xbox 360',
        'user_score']], digits=2)

# combined score averages
threedstotalavg <- round(mean(allgamescopy[allgamescopy$platform == '3DS',
        'total_score']], digits=2)
nin64totalavg <- round(mean(allgamescopy[allgamescopy$platform == 'Nintendo 64',
        'total_score']], digits=2)
ps4totalavg <- round(mean(allgamescopy[allgamescopy$platform == 'PlayStation 4',
        'total_score']], digits=2)
switchtotalavg <- round(mean(allgamescopy[allgamescopy$platform == 'Switch',
        'total_score']], digits=2)
xbox1totalavg <- round(mean(allgamescopy[allgamescopy$platform == 'Xbox One',
        'total_score']], digits=2)
dreamcasttotalavg <- round(mean(allgamescopy[allgamescopy$platform == 'Dreamcast',
        'total_score']], digits=2)
pctotalavg <- round(mean(allgamescopy[allgamescopy$platform == 'PC',
        'total_score']], digits=2)
ps5totalavg <- round(mean(allgamescopy[allgamescopy$platform == 'PlayStation 5',
        'total_score']], digits=2)
wiitotalavg <- round(mean(allgamescopy[allgamescopy$platform == 'Wii',
        'total_score']], digits=2)
xboxsxtotalavg <- round(mean(allgamescopy[allgamescopy$platform == 'Xbox Series X',
        'total_score']], digits=2)
dstotalavg <- round(mean(allgamescopy[allgamescopy$platform == 'DS',
        'total_score']], digits=2)
pstotalavg <- round(mean(allgamescopy[allgamescopy$platform == 'PlayStation',
        'total_score']], digits=2)
psvitatotalavg <- round(mean(allgamescopy[allgamescopy$platform == 'PlayStation Vita',
        'total_score']], digits=2)
wiiutotalavg <- round(mean(allgamescopy[allgamescopy$platform == 'Wii U',
        'total_score']], digits=2)
gbatotalavg <- round(mean(allgamescopy[allgamescopy$platform == 'Game Boy Advance',
        'total_score']], digits=2)
ps2totalavg <- round(mean(allgamescopy[allgamescopy$platform == 'PlayStation 2',
        'total_score']], digits=2)
psptotalavg <- round(mean(allgamescopy[allgamescopy$platform == 'PSP',
        'total_score']], digits=2)
xboxtotalavg <- round(mean(allgamescopy[allgamescopy$platform == 'Xbox',
        'total_score']], digits=2)

```

```

gamecubetotalavg <- round(mean(allgamescopy[allgamescopy$platform == 'GameCube',
                                     'total_score']), digits=2)
ps3totalavg <- round(mean(allgamescopy[allgamescopy$platform == 'PlayStation 3',
                                     'total_score']), digits=2)
stadiatotalavg <- round(mean(allgamescopy[allgamescopy$platform == 'Stadia',
                                     'total_score']), digits=2)
xbox360totalavg <- round(mean(allgamescopy[allgamescopy$platform == 'Xbox 360',
                                     'total_score']), digits=2)

avgcoreplatform = data.frame(platform=c('3DS', 'Nintendo 64', 'PlayStation 4', 'Switch',
                                     'Xbox One', 'Dreamcast', 'PC', 'PlayStation 5',
                                     'Wii', 'Xbox Series X', 'DS', 'PlayStation',
                                     'PlayStation Vita', 'Wii U',
                                     'Game Boy Advance', 'PlayStation 2', 'PSP',
                                     'Xbox', 'GameCube', 'PlayStation 3', 'Stadia',
                                     'Xbox 360'),
                             avg_user_score=c(threedsuseravg, nin64useravg, ps4useravg,
                                     switchuseravg, xbox1useravg, dreamcastuseravg,
                                     pcuseravg, ps5useravg, wiuseravg,
                                     xboxsxuseravg, dsuseravg, psuseravg,
                                     psvitauseravg, wiiumuseravg, gbatotalavg,
                                     ps2totalavg, psptotalavg, xboxtotalavg,
                                     gamecubeuseravg, ps3useravg, stadiuseravg,
                                     xbox360useravg),
                             avg_meta_score=c(threedsmetaavg, nin64metaavg, ps4metaavg,
                                     switchmetaavg, xbox1metaavg, dreamcastmetaavg,
                                     pcmetaavg, ps5metaavg, wiimetaavg,
                                     xboxsxmetaavg, dsmetaavg, psmetaavg,
                                     psvitametaavg, wiiumetaavg, gbametaavg,
                                     ps2metaavg, pspmetaavg, xboxmetaavg,
                                     gamecubemetaavg, ps3metaavg, stadiametaavg,
                                     xbox360metaavg),
                             avg_total_score=c(threedstotalavg, nin64totalavg, ps4totalavg,
                                     switchtotalavg, xbox1totalavg, dreamcasttotalavg,
                                     pctotalavg, ps5totalavg, wiitotalavg,
                                     xboxsxtotalavg, dstotalavg, pstotalavg,
                                     psvitatotalavg, wiitutotalavg, gbatotalavg,
                                     ps2totalavg, psptotalavg, xboxtotalavg,
                                     gamecubetotalavg, ps3totalavg, stadiatotalavg,
                                     xbox360totalavg))

```

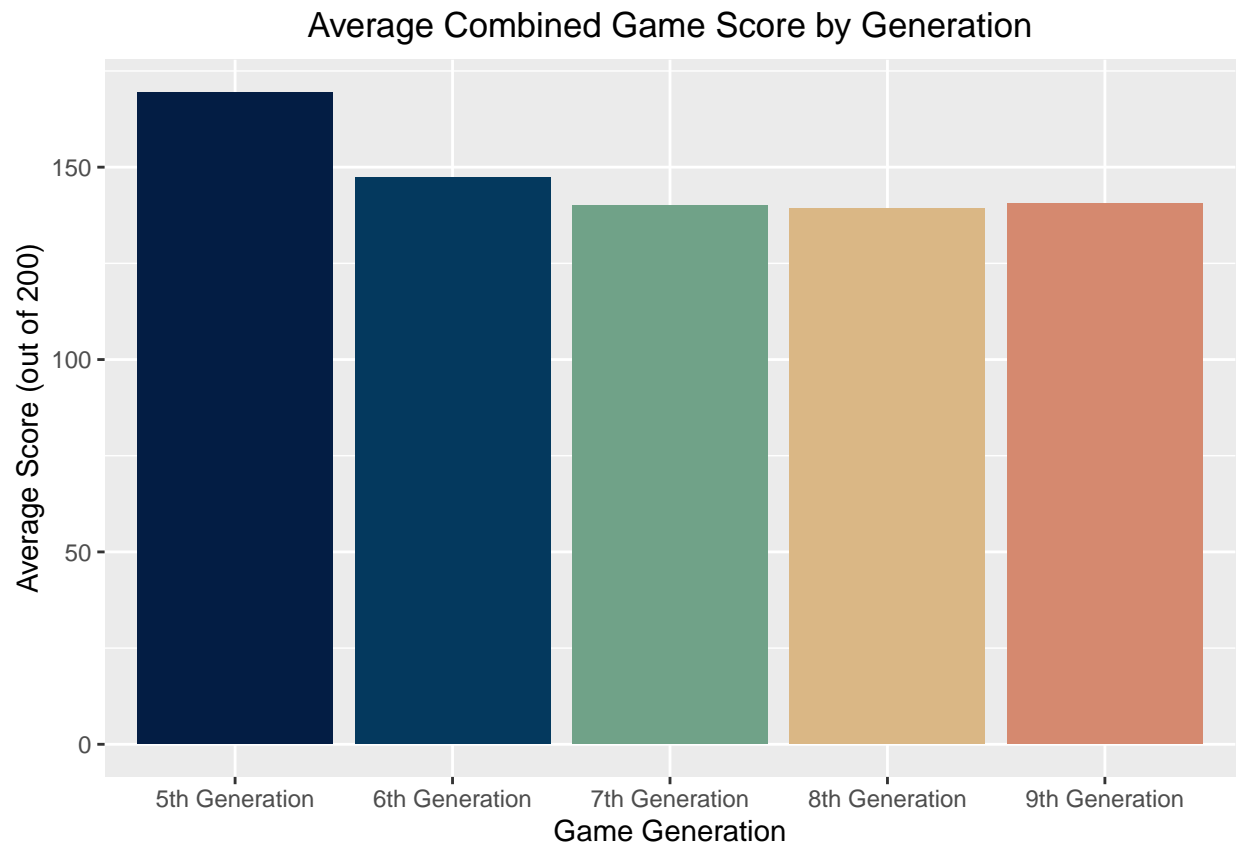
3: Graphing time!

Average Combined Score by Generation

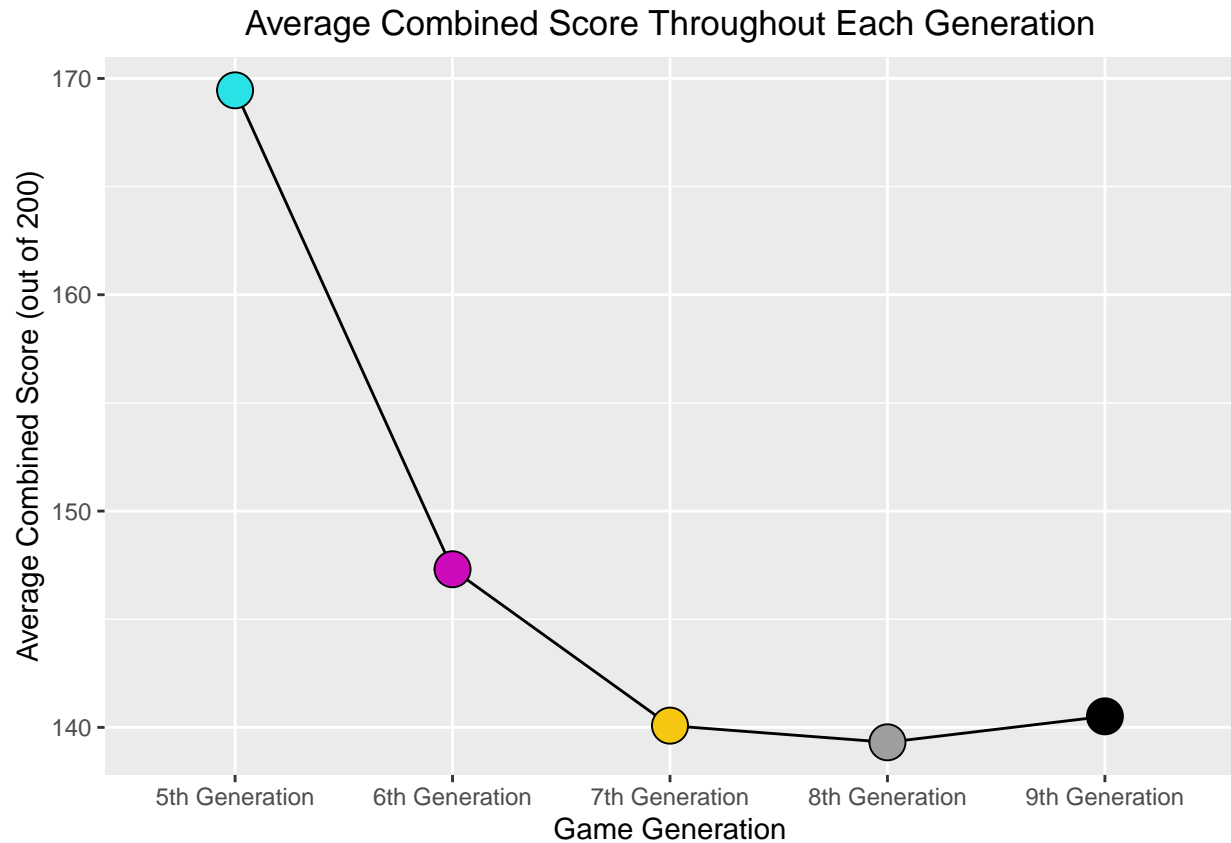
```

ggplot(avgcoregen) + geom_bar(mapping = aes(x=generation, y=avg_total_score,
                                     fill=generation), stat="identity") +
  labs(title="Average Combined Game Score by Generation",
       x="Game Generation", y="Average Score (out of 200)") +
  theme(plot.title=element_text(hjust = 0.5), legend.position="none") +
  scale_fill_manual(values=c("#031D44", "#04395E", "#70A288", "#DAB785", "#D5896F"))

```



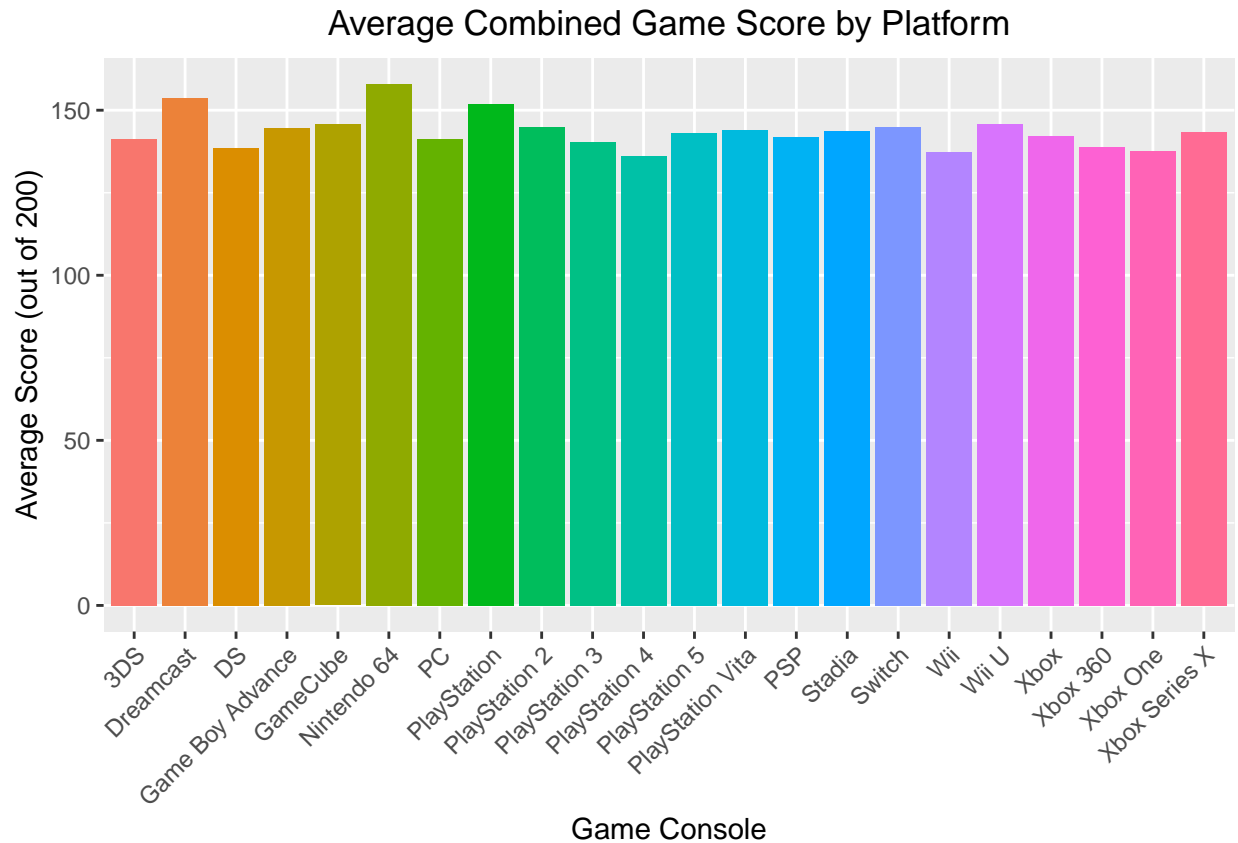
```
ggplot(avgscoregen, aes(x=generation, y=avg_total_score, group=1)) +  
  geom_line() +  
  geom_point(size=6, bg=avgscoregen$generation, pch=21) +  
  labs(title="Average Combined Score Throughout Each Generation", x="Game Generation",  
        y="Average Combined Score (out of 200)") +  
  theme(plot.title=element_text(hjust = 0.5))
```



Average Combined Score by Platform

```
summary(allgamescopy$platform)
```

```
ggplot(avgscoreplatform) + geom_bar(mapping = aes(x=platform, y=avg_total_score,
                                                    fill=platform), stat="identity") +
  labs(title="Average Combined Game Score by Platform",
        x="Game Console", y="Average Score (out of 200)") +
  theme(plot.title=element_text(hjust = 0.5), legend.position="none") +
  theme(axis.text.x = element_text(angle=45, vjust=1, hjust=1))
```



Intermission: Converting wide to long format

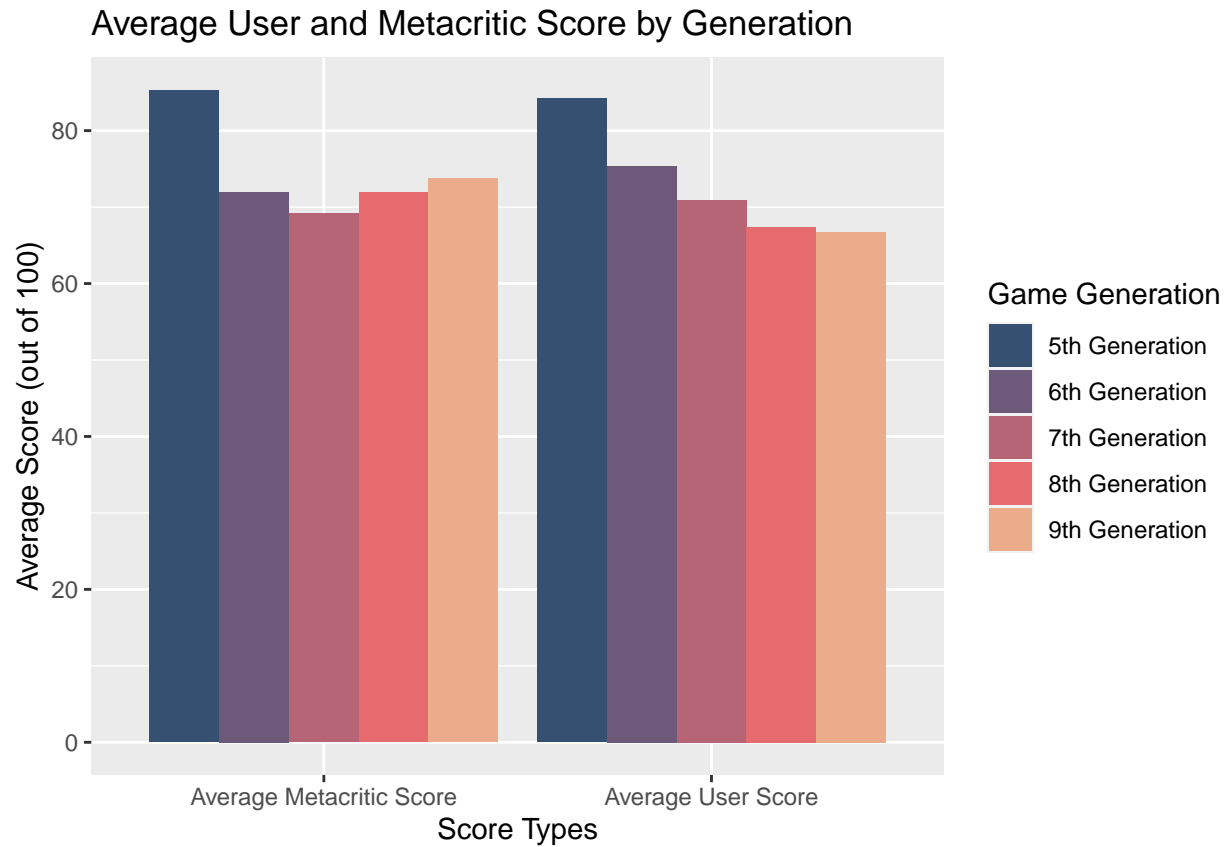
Before I could graph the two columns (user and meta score) together, I needed to put them in long format.

```
avguservsmetagen <- pivot_longer(avgscoregen, cols = c("avg_user_score",
                                                       "avg_meta_score"),
                                names_to = "score_type", values_to = "average_score")

avguservsmetaplat <- pivot_longer(avgscoreplatform, cols = c("avg_user_score",
                                                             "avg_meta_score"),
                                names_to = "platform_type", values_to = "average_score")
```

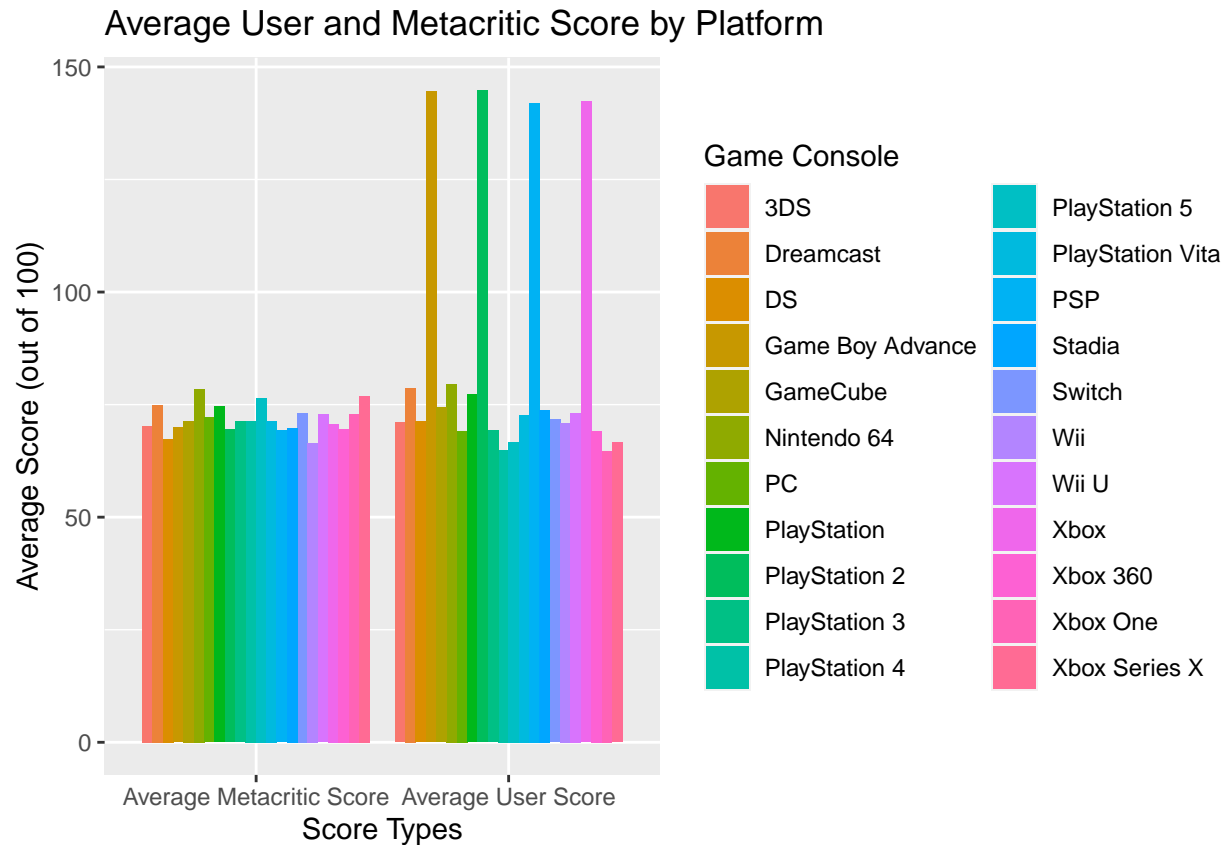
Average Metacritic and User Score by Generation

```
ggplot(avguservsmetagen, aes(fill=generation, y=average_score, x=score_type)) +
  geom_bar(position="dodge", stat="identity") +
  labs(title="Average User and Metacritic Score by Generation", x="Score Types",
       y="Average Score (out of 100)", fill="Game Generation") +
  scale_x_discrete(labels= c("Average Metacritic Score", "Average User Score")) +
  scale_fill_manual(values=c("#355070", "#6D597A", "#B56576", "#E56B6F", "#EAAC8B"))
```



Average Metacritic and User Score by Platform

```
ggplot(avgusersvsmetaplat, aes(fill=platform, y=average_score, x=platform_type)) +
  geom_bar(position="dodge", stat="identity") +
  labs(title="Average User and Metacritic Score by Platform", x="Score Types",
        y="Average Score (out of 100)", fill="Game Console") +
  scale_x_discrete(labels= c("Average Metacritic Score", "Average User Score"))
```



```
ggplot(avgusersvsmetaplat, aes(x=platform, y=average_score, color=platform_type, group=1)) +
  geom_point(size = 6) +
  labs(title="Average User and Metacritic Score by Generation",
        x="Game Console", y="Average Score", color="Score Type") +
  theme(plot.title=element_text(hjust = 0.5)) +
  theme(axis.text.x = element_text(angle=45, vjust=1, hjust=1)) +
  scale_colour_manual(labels=c("Average Metacritic Score", "Average User Score"),
                      values=c("#006e90", "#f18f01"))
```

