Presentation_Workbook

FB Data Challenge Workbook

Here is a rough draft of my final pitch, where I used ggplot2 and tidyverse packages to gain a better understanding of the data I worked with and potential visualization options.

Overall Recommendation

Part A) Movies or TV Shows?

Clean the data.

```
titles_yoy <- read.csv("titles_yoy_growth.csv", na.strings = c("NULL"))
titles_yoy$movie_change[is.na(titles_yoy$movie_change)] <- 0
titles_yoy$tv_change[is.na(titles_yoy$tv_change)] <- 0
titles_yoy
```

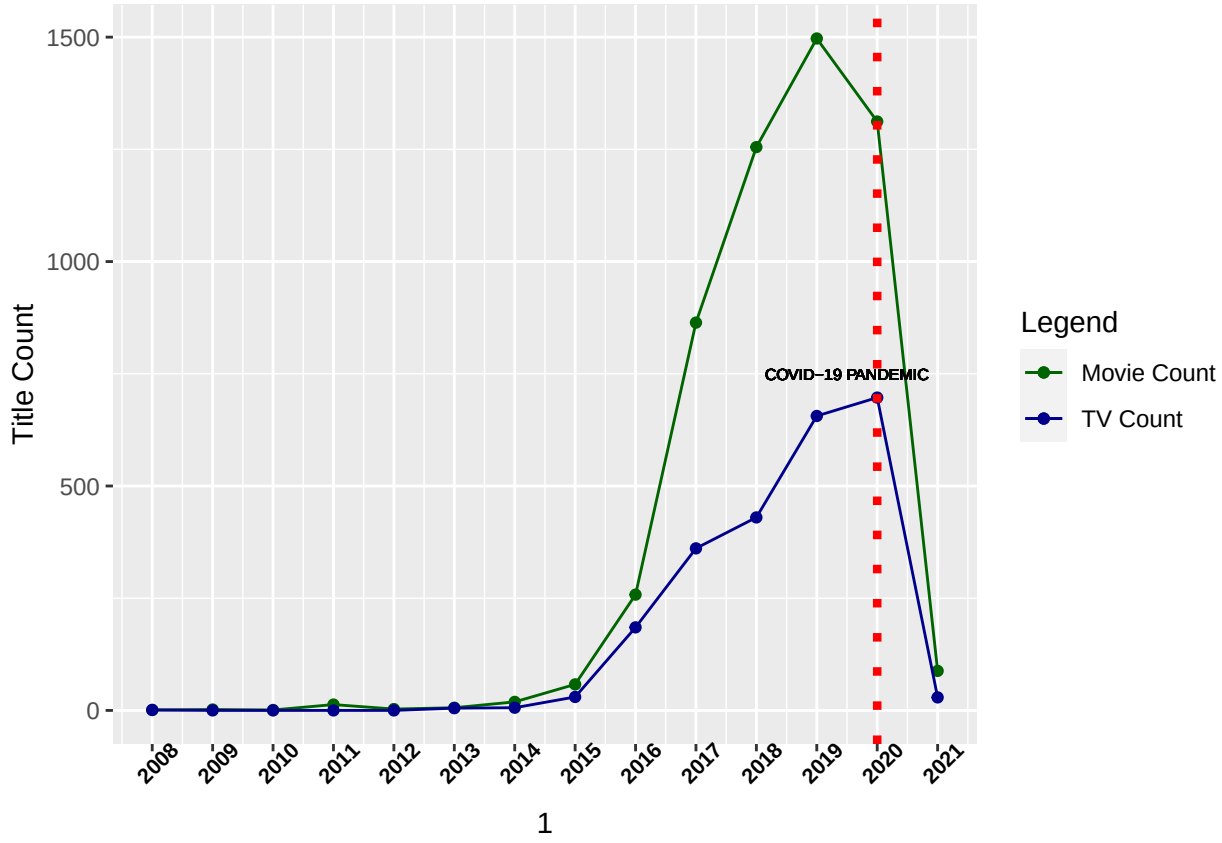
```
##    date_added_year movie_count movie_change tv_count tv_change
## 1             2008           1            0        1         0
## 2             2009           2            1        0        -1
## 3             2010           1           -1        0         0
## 4             2011          13           12        0         0
## 5             2012           3          -10        0         0
## 6             2013           6            3        5         5
## 7             2014          19           13        6         1
## 8             2015          58           39       30        24
## 9             2016         258          200      185       155
## 10            2017         864          606      361       176
## 11            2018        1255          391      430        69
## 12            2019        1497          242      656       226
## 13            2020        1312         -185      697        41
## 14            2021          88        -1224       29      -668
```

```
#write.table(titles_yoy, file = "titles_yoy3.csv", sep = ",", dec = " ", row.names = FALSE)
```

Title Count vs Year Chart

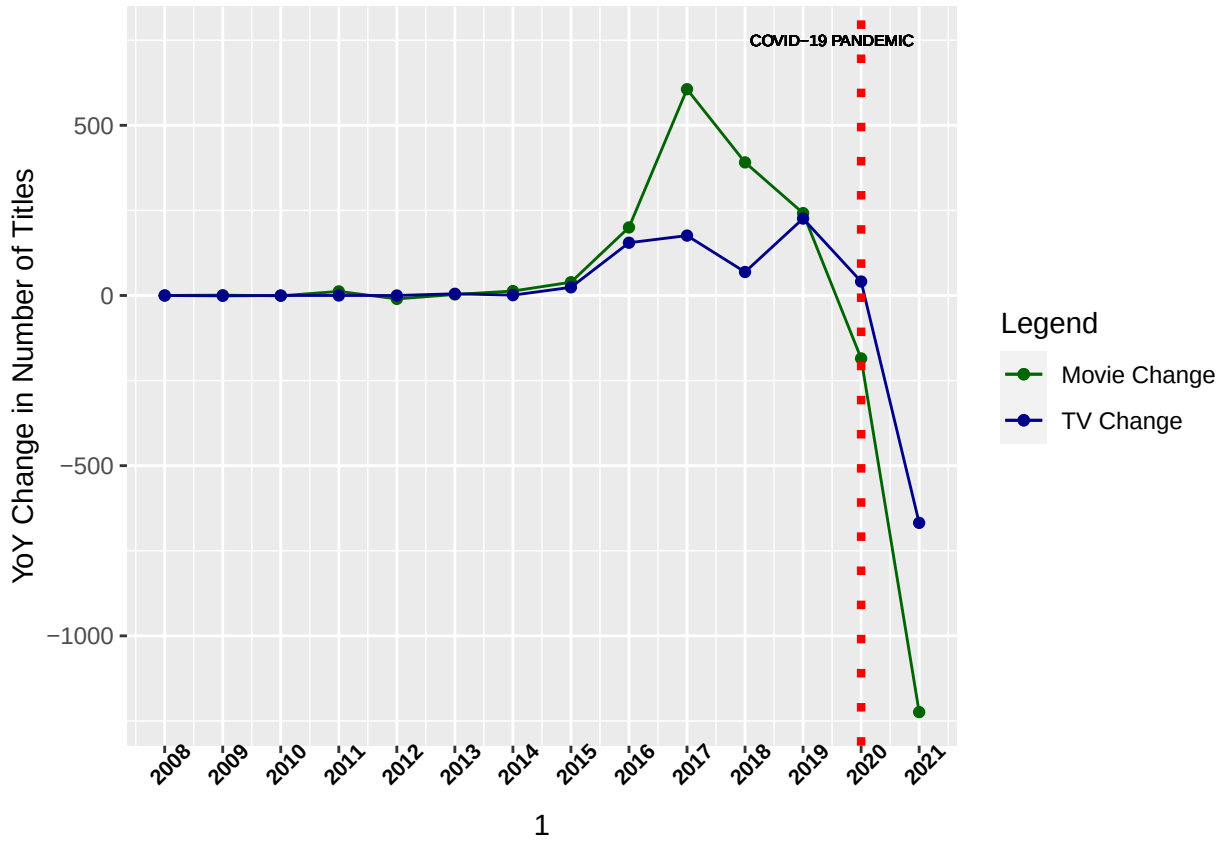
```
colors <- c("Movie Count" = "darkgreen", "TV Count" = "darkblue")
ggplot(data = titles_yoy, aes(x = date_added_year)) +
          geom_line(aes(y = movie_count, color = "Movie Count")) +
          geom_point(aes(y = movie_count, color = "Movie Count")) +
          geom_line(aes(y = tv_count, color = "TV Count")) +
          geom_point(aes(y = tv_count, color = "TV Count")) +
```

```
          labs(x = "Date Added Year", y = "Title Count", color = "Legend") +
          scale_color_manual(values = colors) +
          theme(axis.text.x = element_text(face="bold", color="black",
          size=8, angle=45)) + scale_x_continuous(breaks = seq(2008,2021), 1) +
          geom_vline(xintercept = 2020, linetype = "dotted", size = 1.5, color = "red")+ geom_text(aes
```



YoY Percent Change vs Title Count Chart

```
colors <- c("Movie Change" = "darkgreen", "TV Change" = "darkblue")
ggplot(data = titles_yoy, aes(x = date_added_year)) +
          geom_line(aes(y = movie_change, color = "Movie Change")) +
          geom_point(aes(y = movie_change, color = "Movie Change")) +
          geom_line(aes(y = tv_change, color = "TV Change")) +
          geom_point(aes(y = tv_change, color = "TV Change")) +
          labs(x = "Date Added Year", y = "YoY Change in Number of Titles", color = "Legend") +
          scale_color_manual(values = colors) +
          theme(axis.text.x = element_text(face="bold", color="black",
          size=8, angle=45)) + scale_x_continuous(breaks = seq(2008,2021), 1) +
          geom_vline(xintercept = 2020, linetype = "dotted", size = 1.5, color = "red")+ geom_text(aes
```



Movies have more longevity within the business and seem to have a much higher growth rate from 2015-2019, but started experiencing a decline pre-pandemic. TV Shows were expercincing steady growth during that period as well, but at much lower rates.

There are two ways to interpret these numbers given the basic data we have: a) we can assume the availability of titles correlate to Netflix responding to viewership or user demand b) investment in titles is based off of where Netflix feels it needs to grow in either due to competitors, internal finances, or something else (much of that data is private).

Given Zuckflix's limited budget, the decision is do we want to invest in areas where Netflix/other competitors are already established because we assume viewership is guiding their investment OR do we want to not subject ourselves to that competition and invest in another space but with the knowledge that we will have less of an idea as to what user engagement will look like.

Genre Recommendation: TV Shows

PART B) Genre and User Rating

Prior Genre Analysis

Genre

First, we can look at things in overall counts of what is available on the platform. Here are the top ten genres in terms of largest number of overall title counts (bearing in mind that titles can be counted within multiple categories).

```
genre_count <- read.csv("genre_count.csv")

#ggplot(data = genre_count, aes(x = reorder(genre, count))) + geom_col(aes(y = count), fill = "lightblu
```

User Rating

Now, we can see how what is available on Netflix's platform compares to IMDB user ratings of titles from different genres (keeping in mind that roughly 30% of the totl dataset doesn't have corresponding user ratings, and ~42% of TV shows don't have ratings).

```
genre_user_rating <- read.csv("genre_user_rating.csv")

#ggplot(data = genre_user_rating, aes(x = reorder(genre, avg_user_rating))) + geom_col(aes(y = avg_user
```

Overlap

Follow up question - where is there overlap?

```
inner_join(genre_count,genre_user_rating, by = c("genre" = "genre"))
```

```
##              genre count avg_user_rating
## 1        TV Dramas   704            7.26
## 2   Crime TV Shows   427            7.34
## 3       Docuseries   353            7.40
## 4 British TV Shows   232            7.34
```

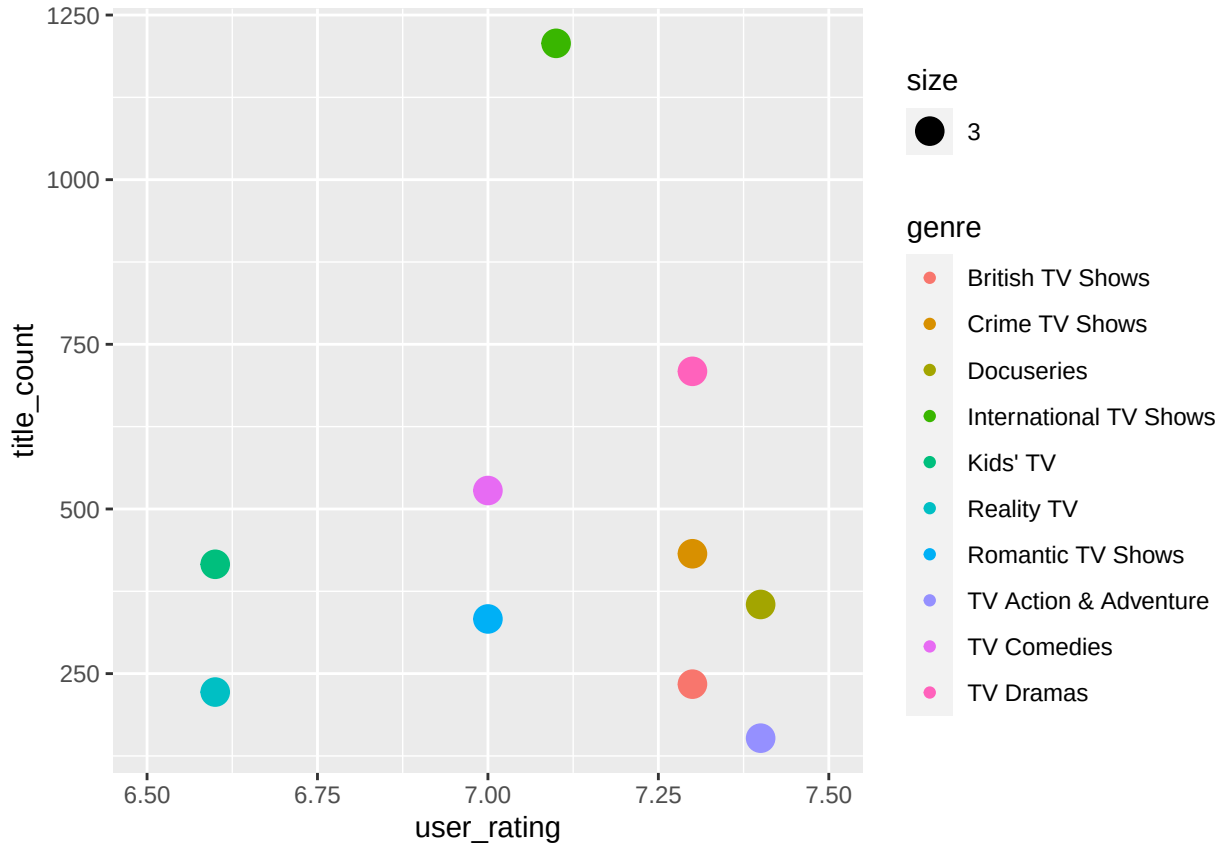
Genre recommendation: Docuseries and TV Dramas

New Genre Analysis

Instead, we can look at how each genre fares in terms of title count and user rating all in the same chart. However, this makes it a bit difficult to determine how much weight to give to each dimension.

```
# scatterplot
scatter_genre <- read.csv("scatter_genre.csv")

ggplot(data = scatter_genre, aes(x=user_rating, y=title_count)) + geom_point(aes(color = genre, size = 3
```



A better way to visualize this in Tableau would be a diverging bar chart, which can be seen in the slide deck.

```
scatter_genre_user <- read.csv("scatter_genre_user.csv")
scatter_genre_user
```

```
##                    genre user_rating title_count
## 1       Classic & Cult TV         8.2          28
## 2            Anime Series         7.5         151
## 3              Docuseries         7.4         355
## 4     Science & Nature TV         7.4          86
## 5  TV Action & Adventure         7.4         152
## 6        British TV Shows         7.3         234
## 7          Crime TV Shows         7.3         432
## 8               TV Dramas         7.3         709
## 9     TV Sci-Fi & Fantasy         7.3          78
## 10           TV Thrillers         7.2          51
```

```
# diverging bar chart
inner_join(scatter_genre, scatter_genre_user, by = c("genre" = "genre", "user_rating" = "user_rating",
```

```
##                    genre title_count user_rating
## 1              TV Dramas         709         7.3
## 2          Crime TV Shows         432         7.3
## 3             Docuseries         355         7.4
## 4        British TV Shows         234         7.3
## 5 TV Action & Adventure         152         7.4
```

Part C) Maturity Rating and Duration

This part was largely edited in Tableau.

Maturity Rating Recommendation: TV-MA or TV-14

Duration

In order to quickly determine the number of seasons we should recommend per title, we can look at the average number of seasons per genre-rating combination. From this we can suggest the initial anticipate duration of the title should be between 1-2 seasons.

```
duration_genre <- read.csv("duration_genre.csv")
duration_genre
```

```
##       genre_rating avg_num_seasons
## 1 Docuseries TV-14            1.22
## 2 Docuseries TV-MA            1.45
## 3  TV Dramas TV-14            1.96
## 4  TV Dramas TV-MA            1.77
```

Duration Recommendation: 1-2 seasons

Country Specific Recommendation

Choosing a Country

What are we optimizing for?

1. *Facebook's main source of revenue: ads.* "... while nearly 60% of its daily active users live outside North America and Europe, those users only account for about one-quarter of Facebook's total revenues". Thus, 'Zuckflix' should be looking to invest in an emerging market outside of the West that has the potential to monetize content, i.e. link 'ZuckFlix' titles to existing FB products and ad revenue.
2. India is one of the largest emerging markets according to Goldman Sach's 2021 EME Report. "... The increased digitalization and internet adoption is *benefitting new-age companies operating in the fields of technology*, communication services and online services.
3. And, "*Southern Asia has over 1.126 billion users that are still unconnected*, making it a top emerging market." Given Facebook's (and other tech companies) shift to investing in low bandwith markets, India is a prime country to look at.
4. *India has Facebook's largest existing user base* (330 mil), meaning it is already established within in the region which will facilitate marketing and connections to ad revenue.
5. *India is not within Netflix's [top ten leading markets] (https://www.statista.com/statistics/499844/netflix-markets-penetration/)* (by number of subscribers) which decreases competition, although Amazon Prime Video has had increasingly high market penetration in the region.

Quantifying India's Market Potential

Mobile vs Broadband

While we know that India is a top emerging market, we need to understand the existing landscape in order to determine the best market entry strategy. To start, we can look at broadband vs mobile usage over the years to determine how our users will access our product. source

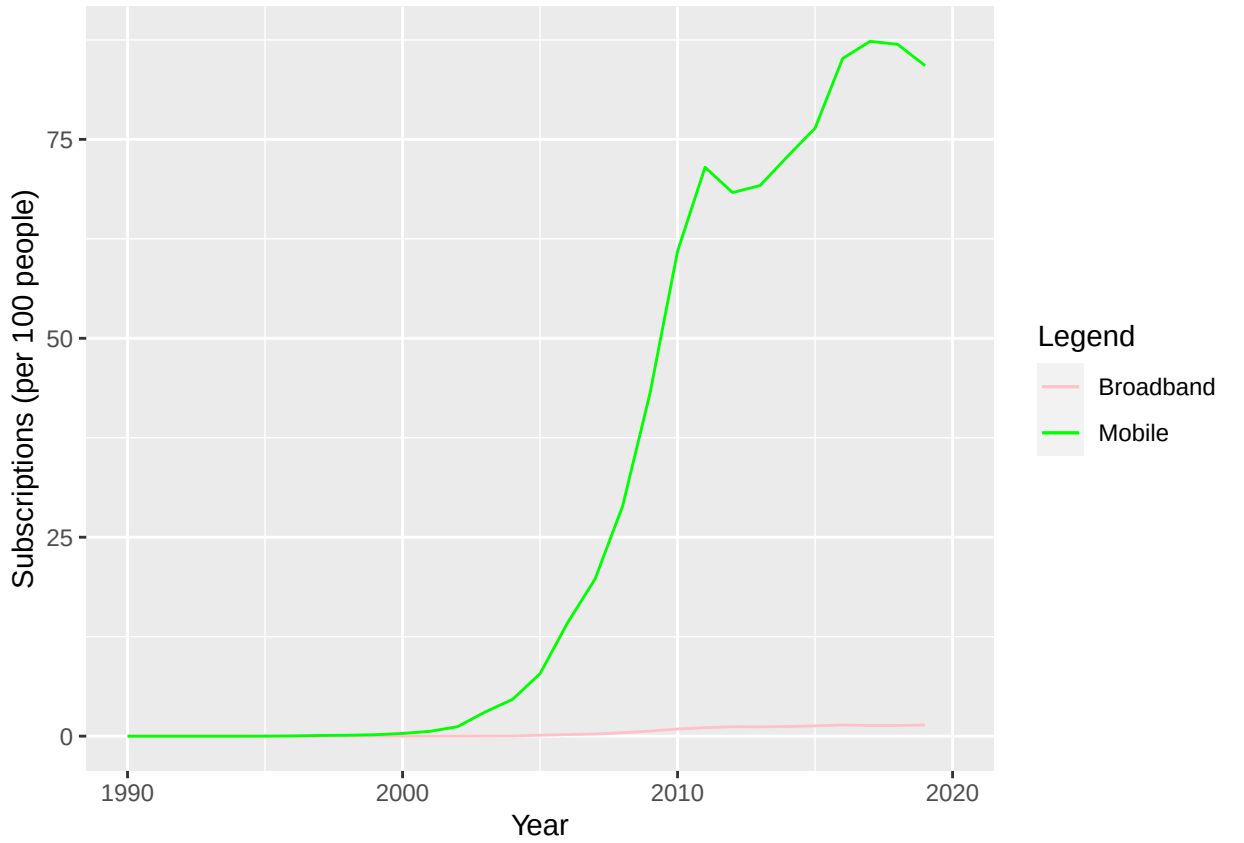
```
# broadband subscriptions (per 100 ppl)
broadband <- read.csv("broadband_sub.csv")
broadband <- broadband %>% filter(Country.Name == "India")
broadband <- broadband %>% pivot_longer(cols = starts_with("X"), names_to = "year", values_to = "broadba
broadband <- broadband[,5:6]
broadband$year <- substring(broadband$year, 2, length(broadband$year))
broadband$year <- as.integer(broadband$year)

# mobile subscriptions (per 100 ppl)
mobile <- read.csv("mobile_subww.csv")
mobile <- mobile %>% filter(Country.Name == "India")
mobile <- mobile %>% pivot_longer(cols = starts_with("X"), names_to = "year", values_to = "mobile_sub")
mobile <- mobile[,5:6]
mobile$year <- substring(mobile$year, 2, length(mobile$year))
mobile$year <- as.integer(mobile$year)

# look at both overlayed
broad_mobile <- inner_join(broadband, mobile, by = c("year" = "year"))
broad_mobile$broadband_sub[is.na(broad_mobile$broadband_sub)] <- 0
broad_mobile$mobile_sub[is.na(broad_mobile$mobile_sub)] <- 0
broad_mobile <- broad_mobile[-nrow(broad_mobile),]

# write to csv
#write.table(broad_mobile, file = "broad_mobile_tableau.csv", sep = ",", dec = " ", row.names = FALSE)

# graph
colors <- c("Broadband" = "pink", "Mobile" = "green")
ggplot(data = broad_mobile, aes(x=year)) + geom_line(aes(y=broadband_sub, color = "Broadband")) + geom_
```



We can see that mobile subscriptions far outnumber broadband subscriptions, which means we should be optimizing for content that will perform best on smaller screens and mobile devices.

User Access Touchpoint: Mobile

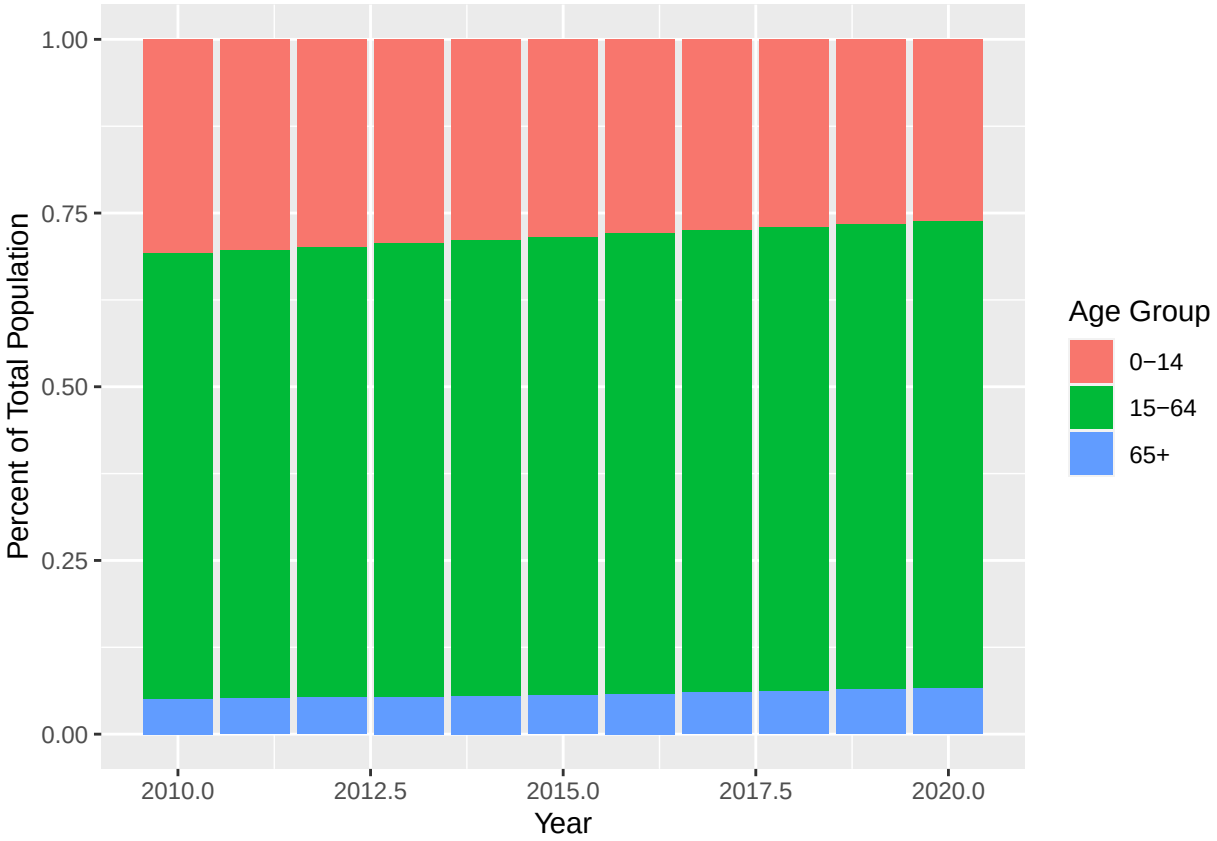
India Age Demographics

Next, let's take a look at India's age demographic breakdown to determine what content rating level would best suit the largest audience. *source1* source2

```
india_age <- read.csv("india_age_breakdown.csv")
colnames(india_age) <- c("year", "0-14", "15-64", "65+")
india_age <- pivot_longer(india_age, cols = c("0-14", "15-64", "65+"), names_to = "age_group", values_t
#india_age$population_perc <- paste(india_age$population_perc, "%")

#write.table(india_age, file = "india_age_tableau_3.csv", sep = ",", dec = ".", row.names = FALSE)


ggplot(data = india_age, aes(x = year, y = population_perc, fill = age_group)) +geom_bar(position = "fi
```



Content Rating Recommendation: Adult Content (i.e. TV-14/PG-13 and up)

Language Breakdown

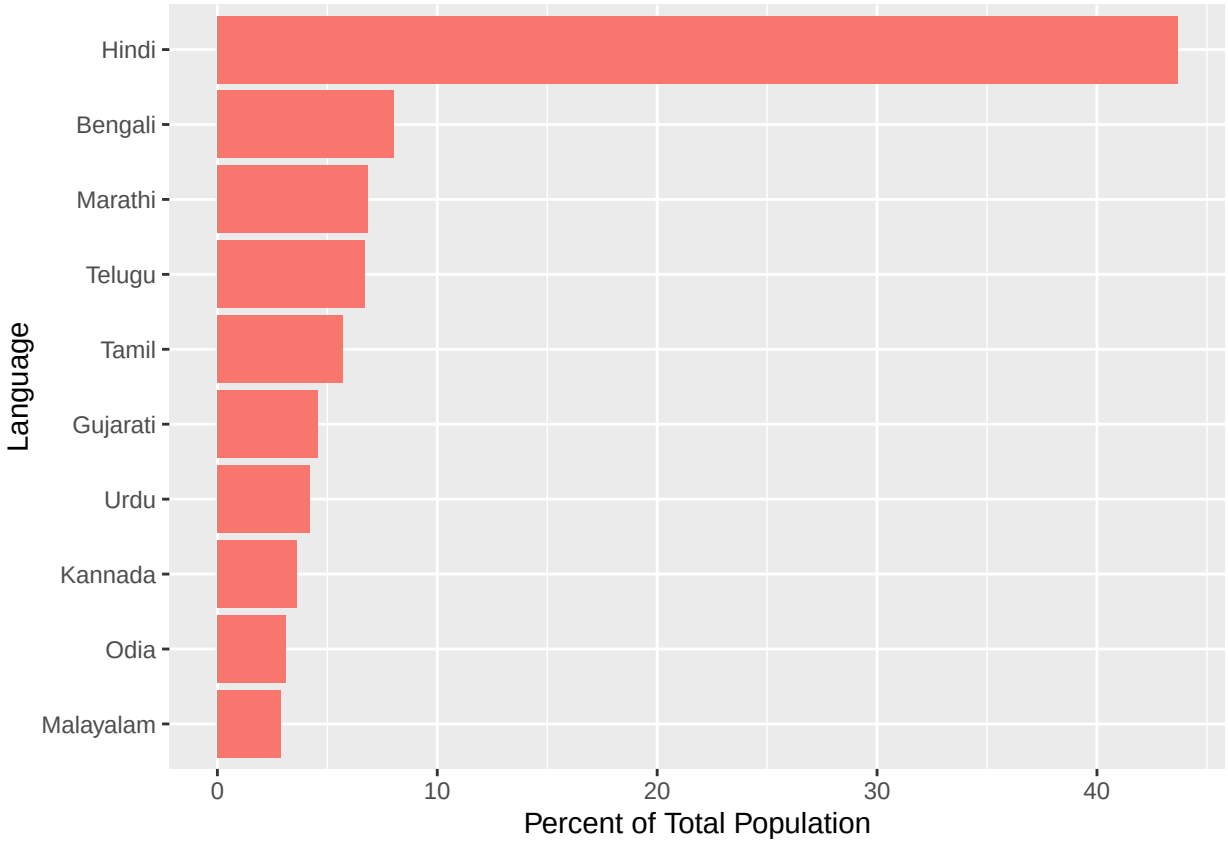
Because India is a multilingual country, content reusability across a vareity of different langauges and cultures will be important; namely, the cost of adding a subtitle is lower than the cost of voice-overs or reproducing the show entirely, so we should produce the original content in such a way that will cater to the widest audience.

```
languages <- read.csv("india_language.csv")
languages <- languages %>% top_n(10)
```

```
## Selecting by Percent
```

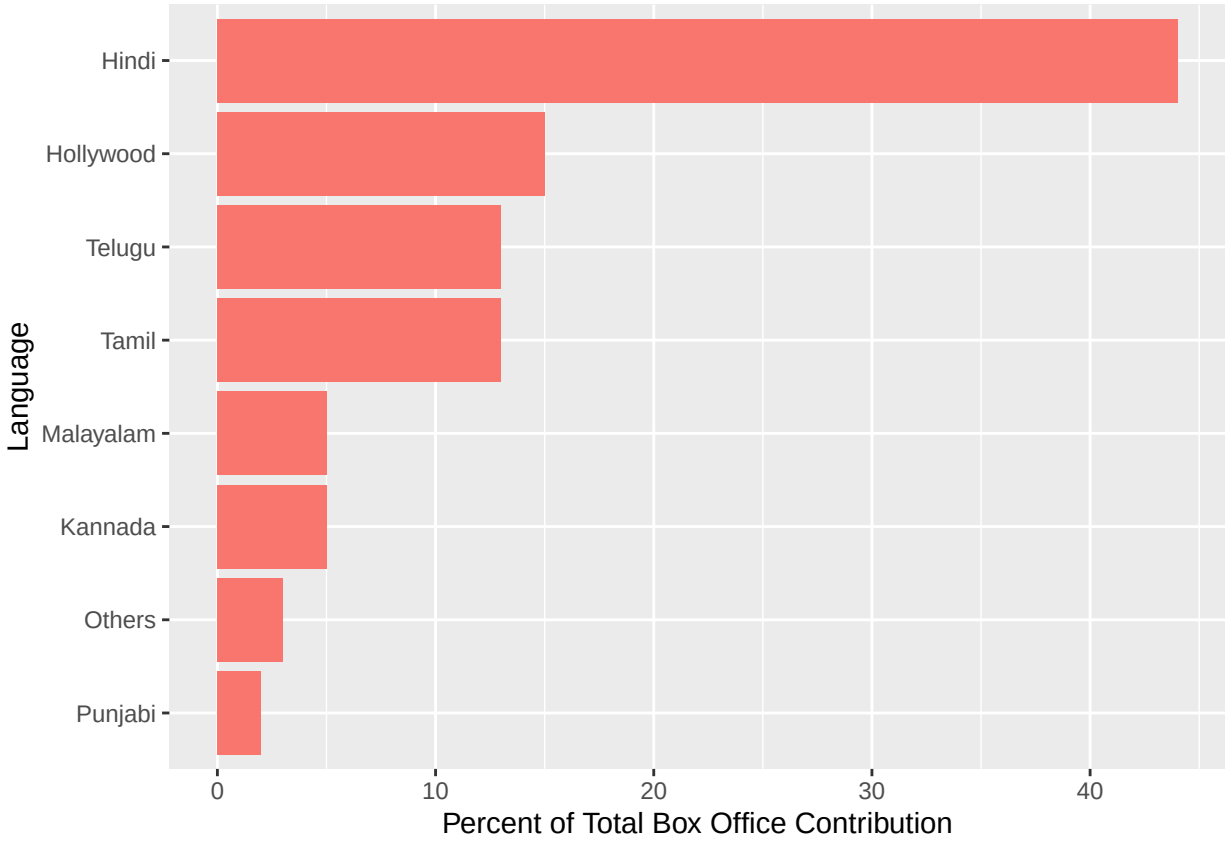
```
#write.table(languages, file = "languages_tableau.csv", sep = ",", dec = " ", row.names = FALSE)

ggplot(data = languages, aes(x=reorder(Language, Percent))) + geom_col(aes(y=Percent, fill = "orange"))
```



Alternatively, we can rank languages in terms of their respective film industry's box office contributions.

```
box_office <- read.csv("box_officed.csv")
ggplot(data = box_office, aes(x=reorder(Language, Percent))) + geom_col(aes(y=Percent, fill = "orange"))
```




```
write.table(box_office, file = "box_office.csv", sep = ",", dec = ".", row.names = FALSE)
```

```
india_combined <- inner_join(languages, box_office, by = c("Language" = "Language"))
india_combined$Percent.x <- round(india_combined$Percent.x,0)
colnames(india_combined) <- c("Language", "Total Population", "Percent of Total Population", "Percent o
india_combined
```

```
##    Language Total Population Percent of Total Population
## 1     Hindi        528347193                          44
## 2    Telugu         81127740                           7
## 3     Tamil         69026881                           6
## 4   Kannada         43706512                           4
## 5 Malayalam         34838819                           3
##   Percent of Box Office Contribution NA
## 1                                 44 NA
## 2                                 13 NA
## 3                                 13 NA
## 4                                  5 NA
## 5                                  5 NA
```

```
write.table(india_combined, file = "india_combined.csv", sep = ",", dec = ".", row.names = FALSE)
```

Language Recommendation: Hindi (Bollywood

Content Recommendation

Movie or TV Show?

There are 990 distinct titles.

Let's look at the title breakdown by title type. Unsurprisingly, we see that there are more movies rather than TV shows, which makes sense given that the Bollywood film industry is worth over a couple billion dollars and India produces the most films out of any country.

```
title_type <- read.csv("india_titletype.csv")
title_type
```

```
##   num_movies num_shows
## 1        915        75
```

Genre

We can look at genres filtered by country and the above maturity rating. It is important to note that in this dataset, Indian movies are filtered by American TV status (TV-14, TV-MA) as well as Movie status (PG-13, R) so both ratings are taken into account when conducting this genre-based analysis.

This part was largely done in Tableau.

Genre Recommendation: Comedies and Dramas (Indian staples!)