Algoritma Academy: Data Visualization

Samuel Chan

13 September, 2024

Before you go ahead and run the codes in this coursebook, it's often a good idea to go through some initial setup. Under the *Libraries and Setup* tab you'll see some code to initialize our workspace, and the libraries we'll be using for the projects. You may want to make sure that the libraries are installed beforehand by referring back to the packages listed here. Under the *Training Focus* tab we'll outline the syllabus, identify the key objectives and set up expectations for each module.

Background

Algoritma

The following coursebook is produced by the team at Algoritma for its Data Science Academy workshops. The coursebook is intended for a restricted audience only, i.e. the individuals and organizations having received this coursebook directly from the training organization. It may not be reproduced, distributed, translated or adapted in any form outside these individuals and organizations without permission.

Algoritma is a data science education center with bootcamp programs offered in:

- Bahasa Indonesia (Jakarta campus)
- English (Singapore campus)

Lifelong Learning Benefits

If you're an active student or an alumni member, you also qualify for all our future workshops, 100% free of charge as part of your **lifelong learning benefits**. It is a new initiative to help you gain mastery and advance your knowledge in the field of data visualization, machine learning, computer vision, natural language processing (NLP) and other sub-fields of data science. All workshops conducted by us (from 1-day to 5-day series) are available to you free-of-charge, and the benefits **never expire**.

Second Edition

This coursebook is initially written in 2017.

This is the second edition, written in late August 2020. Some of the code has been refactored to work with the latest major version of R, version 4.0. I would like to thank the incredible instructor team at Algoritma for their thorough input and assistance in the authoring and reviewing process.

Libraries and Setup

We'll set-up caching for this notebook given how computationally expensive some of the code we will write can get.

```
options(scipen = 9999)
rm(list=ls())
knitr::opts_chunk$set(
  message = FALSE,
  warning = FALSE
)
```

You will need to use `install.packages()` to install any packages that are not already downloaded onto your machine. You then load the package into your workspace using the `library()` function:

```
library(ggplot2)
library(GGally)
library(ggthemes)
library(ggpubr)
library(leaflet)
library(lubridate)
```

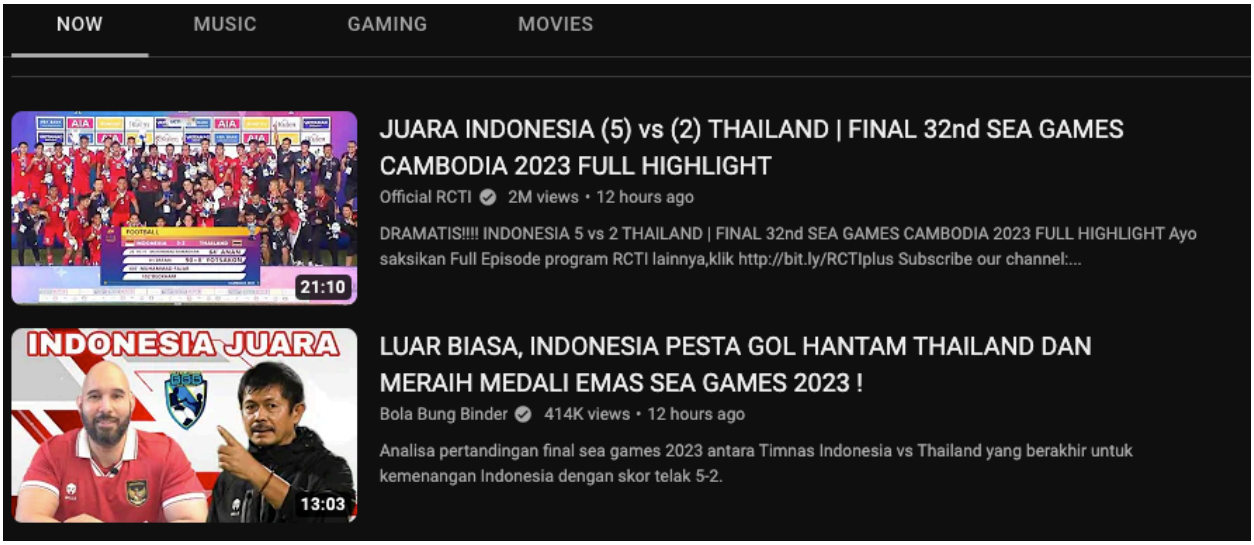
The data we'll be working on is a rather recent Youtube Trending Videos dataset[1]. It has 36.800 records of trending videos between 1st July 2022 to 31st December 2022, and on each record of trending video is a list of variables:

General information relating to video
- Trending date
- Title (video title)
- Channel Title
- Category ID
- Publish Time
- Comments (Disabled?)
- Ratings (Disabled?)
- Video error or removed?

Statistics on a particular date - Views
- Likes
- Dislikes
- Comment Count

Illustration of a "Trending" section on YouTube:

[1]This dataset is first contributed under CC0 public domain by Mitchell J (datasnaek on Kaggle), and maintained by other contributors. It has 13,400 records of trending videos between 14th November 2017 to 21st January 2018.



Training Objectives

The primary objective of this course is to provide a fun and hands-on session to help participants gain full proficiency in data visualization systems and tools. You will learn to create compelling narratives by combining charting elements with a rich grammar under the guidance of the lead instructor and our team of teaching assistants.

- **Plotting Essential**
- Revision: Built-in Plotting Functionalities
- Revision: Scatterplot, Histogram, Line and Column Bars
- Axis, Title and Panel Styles
- Grammar of Graphics
- `ggplot2` Basics
- **Plotting Better**
- Using Themes
- Multi-dimensional Faceting
- Visualizing Geo-Spatial Data with `leaflet`
- Lattice Plotting system

By the end of the workshop, Academy students can choose to complete either of the Learn-By-Building modules as their graded assignment:

Ready for Publication
Applying what you've learned, create a visualization that is polished with the appropriate annotations, aesthetics and some simple commentary. This can be any visualization using the YouTube dataset, but it should communicate a story.

Interactive Map
Create a web page with an interactive map embedded on it. Use a custom icon for the map markers to represent business locations, and show details about each location pin ("markers") upon user's interaction.

This graded assignment is worth **(2) Points**.

Plotting Essentials

R as a statistical computing environment packs a generous amount of tools allowing us to reshape, clean and visualize our data through its built-in capabilities. In the first part of this coursebook, we'll take a look at many of these capabilities and learn how to incorporate these into our day-to-day data science work.

In the second part of this coursebook, we'll shift our focus onto `ggplot`, a plotting system by Hadley Wickham. As you'll see in this 3 days workshop, this plotting system is among the most popular visualization tools today because of its power, extensibility and simplicity (an unlikely combination).

To get started with plotting in R, let's start by reading our data into the environment:

```
vids <- read.csv("data_input/USvideos_2023.csv")
names(vids)
```

```
##  [1] "trending_date"          "title"                 "channel_title"
##  [4] "category_id"            "publish_time"          "views"
##  [7] "likes"                  "dislikes"              "comment_count"
## [10] "comments_disabled"      "ratings_disabled"      "video_error_or_removed"
```

Taking a quick peek at the `trending_date` column reveals that the date values are stored in a year-day-month format, with `.` (dot) being the delimiter:

```
head(vids$trending_date)
```

```
## [1] "23.01.01" "23.01.01" "23.01.01" "23.01.01" "23.01.01" "23.01.01"
```

Because these values follows the `yy.dd.mm` format, our data processing steps would handle this format accordingly through the `format` argument:

```
vids$trending_date <- as.Date(vids$trending_date, format="%y.%d.%m")
```

In real life, most date formats uses a combination that maps to the following:

- `%Y`: 4-digit year (1982)
- `%y`: 2-digit year (82)

- `%m`: 2-digit month (01)
- `%d`: 2-digit day of the month (13)
- `%A`: weekday (Wednesday)
- `%a`: abbreviated weekday (Wed)
- `%B`: month (January)
- `%b`: abbreviated month (Jan)

You should follow your lead instructor on a few in-classroom practice to increase your familiarity with the `as.Date()` function, as processing dates are a fairly common process in many data analysis routines:

```
# Demo:
as.Date("Aug 30,2020", format = "%b %d,%Y")

# Dive Deeper (complete the following):
as.Date("30aug20", format = ___)
as.Date("2020-08-30", format = ___)
as.Date("08.30.2020", format = ___)
```

The raw dataset does not have the proper names for each category, but identify them by an "id" instead. The following code chunk "switches" them by "id" and also convert that to a factor. We will also convert our video titles to a character vector:

```
vids$category_id <- sapply(as.character(vids$category_id), switch,
                           "1" = "Film and Animation",
                           "2" = "Autos and Vehicles",
                           "10" = "Music",
                           "15" = "Pets and Animals",
                           "17" = "Sports",
                           "19" = "Travel and Events",
                           "20" = "Gaming",
                           "22" = "People and Blogs",
                           "23" = "Comedy",
                           "24" = "Entertainment",
                           "25" = "News and Politics",
                           "26" = "Howto and Style",
                           "27" = "Education",
                           "28" = "Science and Technology",
                           "29" = "Nonprofit and Activism",
                           "43" = "Shows")

vids$category_id <- as.factor(vids$category_id)
```

And with that, the next thing we'll need to do is to convert the `publish_time` variable into a date-time class object. Because we're analyzing trending YouTube videos in the US, it makes sense then that we uses a timezone like New York for our analysis:

```
as.POSIXct(head(vids$publish_time),
           format="%Y-%m-%dT%H:%M:%S",
           tz="America/New_York")
```

```
## [1] "2023-01-01 05:12:55 EST" "2022-12-31 20:00:04 EST"
## [3] "2023-01-01 01:11:58 EST" "2022-12-31 14:59:56 EST"
## [5] "2023-01-01 05:21:44 EST" "2022-12-31 20:51:09 EST"
```

Once we've done a sanity check, we can go ahead and apply the conversion to the whole column:

```
vids$publish_time <- as.POSIXct(vids$publish_time,
           format="%Y-%m-%dT%H:%M:%S",
           tz="America/New_York")
```

By this point you may be wondering if this could have been made simpler. Allow me to introduce a popular package by the name of `lubridate`. Using `lubridate`, instead of manually specifying the format, you do it "declaratively", like the following:

```
date <-  c("20.30.8", "20.31.8")
date2 <- c("30-8-20", "31-8-20")
date3 <- c("2017-11-13 17:13:01 EDT")

# base R
as.Date(date, format="%y.%d.%m")
```

```
## [1] "2020-08-30" "2020-08-31"
```

```
as.Date(date2, format="%d-%m-%y")
```

```
## [1] "2020-08-30" "2020-08-31"
```

```
as.POSIXct(date3, format="%Y-%m-%d %H:%M:%S", tz="UTC")
```

```
## [1] "2017-11-13 17:13:01 UTC"
```

```
# lubridate equivalent:
ydm(date)
```

```
## [1] "2020-08-30" "2020-08-31"
```

```
dmy(date2)
```

```
## [1] "2020-08-30" "2020-08-31"
```

```
ymd_hms(date3, tz="UTC")
```

```
## [1] "2017-11-13 17:13:01 UTC"
```

Observe how simple `lubridate` work with dates and time. In fact, when we use `ymd`, `ymd_hms` or one of its variants, these functions recognize the patterns and will identify the right separators as long as the order of formats is correct. These functions will also parse dates correctly even when the input contain differently formatted dates!

Let's see a few more things that `lubridate` can do. I'll subset the data for the most popular trending video (by number of views) and we'll extract information from the `trending_date` of the most popular trending video:

```
most <- vids[vids$views == max(vids$views),]
year(most$trending_date)
```

```
## [1] 2023
```

```
month(most$trending_date)
```

```
## [1] 4
```

```
day(most$trending_date)
```

```
## [1] 29
```

We will also go ahead and create three new variables for our data frame, storing the hours, period of the day, and the day of the week of each video at the time of publish:

```
vids$publish_hour <- hour(vids$publish_time)

pw <- function(x){
    if(x < 8){
      x <- "12am to 8am"
    }else if(x >= 8 & x < 16){
      x <- "8am to 3pm"
    }else{
      x <- "3pm to 12am"
    }
}

vids$publish_when <- as.factor(sapply(vids$publish_hour, pw))
vids$publish_wday <- as.factor(weekdays(vids$publish_time))
```

While the `publish_wday` is now a factor, we can also arrange it so our plots will display them in our desired order:

```
vids$publish_wday <- ordered(vids$publish_wday, levels=c("Monday", "Tuesday", "Wednesday", "Thursday",
```

We'll also go ahead and convert some of the variables into numeric variables as and where appropriate:

```
vids[,c("views", "likes", "dislikes", "comment_count")] <- lapply(vids[,c("views", "likes", "dislikes",
```

Hopefully up to this point, none of the above data transformation and cleansing process looks too unfamiliar for you! If you do need a refresher, refer to the Programming for Data Science coursebook - we're really applying many of the same ideas to a new dataset, and so you should feel somewhat comfortable up to this point of the course :)

`vids` has 13400 records of trending videos, but there are many videos that were trending for a few days and we really only have a collection of 2,986 unique videos. On a very broad average, each video was trending for ~4.5 days.

Let's create a dataframe, call it `vids.u` that takes only the first observation of each `vids.title` within the data. `match` returns a vector of the positions of matches of its first argument in its second:

```
vids.u <- vids[match(unique(vids$title), vids$title),]
```

We'll also create one more variable `timetotrend` to measure the time it takes for a video to become "trending":

```
vids.u$timetotrend <- vids.u$trending_date - as.Date(vids.u$publish_time)
vids.u$timetotrend <- as.factor(ifelse(vids.u$timetotrend <= 7, vids.u$timetotrend, "8+"))
```

The `ifelse()` command is a very quick way to perform a conditional check, and then do one of two things depending if the conditionals evaluate to TRUE or FALSE. Here's another example to illustrate this concept better:

```
hours <- head(vids$publish_hour)
print(hours)
```

```
## [1]  5 20  1 14  5 20
```

```
ifelse(hours >= 18, "night_shift", "day_shift")
```

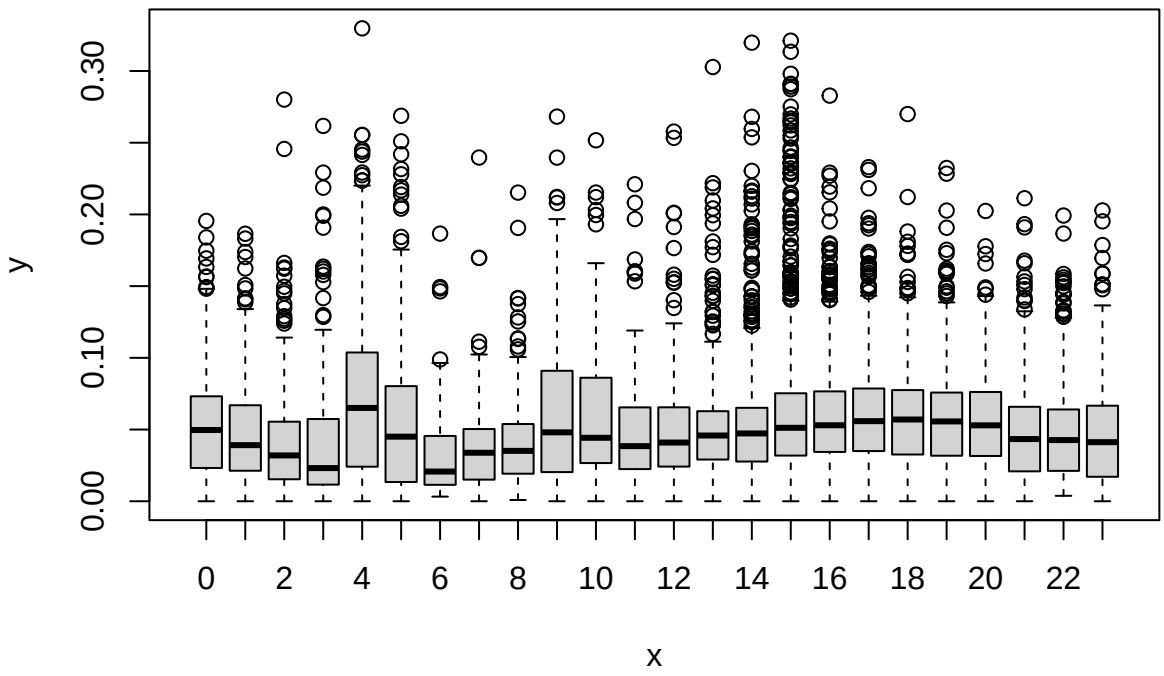
```
## [1] "day_shift"   "night_shift" "day_shift"   "day_shift"   "day_shift"
## [6] "night_shift"
```

With these done, we'll move into the exciting part of this workshop: plotting!

Base Plotting and Statistical Plots

Statistical plots helps us visually inspect our dataset and there are numerous ways to achieve that in R. The simplest of which is through the `plot()` function. In the following code we create two vectors, `x` and `y`, and created a plot:

```
plot(as.factor(vids.u$publish_hour), vids.u$likes/vids.u$views)
```



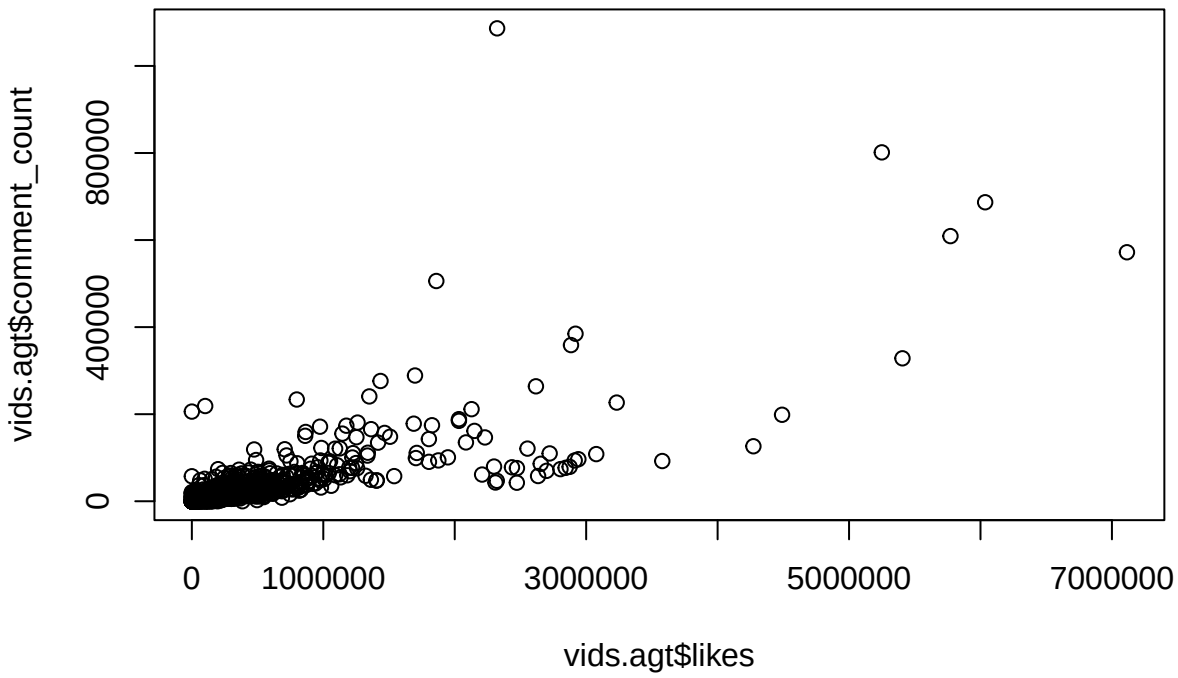
The above gives us a boxplot that compares the "likes ratio" across different period of the day. We want to observe if there is any correlation between the likes-to-view percentage and the period of time when the video was published. As expected, we did not find any obvious patterns, in part because this compares the "likes" a video have on the first day of it being "trending" and the hour when it was published onto YouTube. In a sense, any kind of effect the hour variable can have has been adjusted for (or significantly reduced) by the noise between these two events.

`plot()` knows how to pick sensible defaults based on the input vector it was given. To illustrate this point, I'll subset the data to take only trending videos within the Music, Gaming and Entertainment (every women's favorite 3 things):

```
vids.agt <- vids.u[vids.u$category_id == "Music" |
                     vids.u$category_id == "Gaming" |
                     vids.u$category_id == "Entertainment", ]
```

And notice that as we call `plot` now with two numeric variables, so it creates a scatterplot for us (instead of the boxplot, as seen earlier):

```
plot(vids.agt$likes, vids.agt$comment_count)
```



We'll drop the empty levels from our `category_id` variable, and also create two new variables that measure the likes and comment per video view for each observation:

```
vids.agt$category_id <- factor(vids.agt$category_id)
vids.agt$likesp <- vids.agt$likes/vids.agt$views
vids.agt$commentp <- vids.agt$comment_count/vids.agt$views
```

Our earlier scatter plot really isn't very informative or even pleasant to look at. The key, as it is with data visualization in general, is to have our plot be effective. A plot that is effective complements how human visual perception works. The scatterplot is ineffective because it could be communicating more with less "visual clutter" at the bottom left.

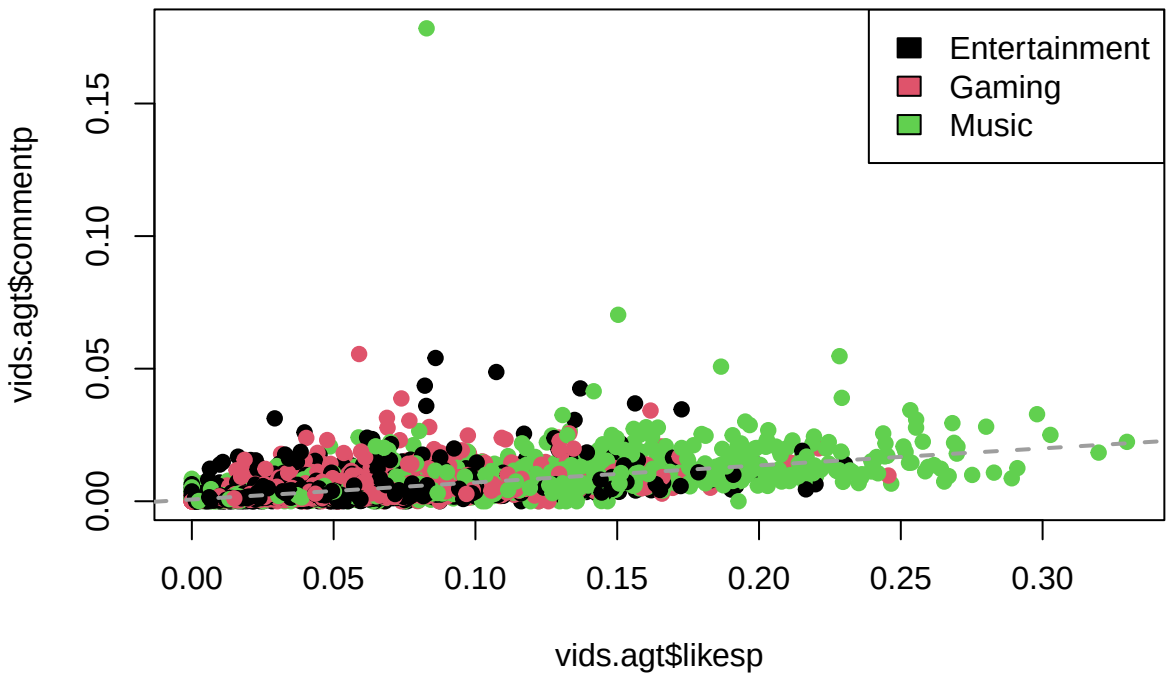
An approach to fix that is by coloring the plots. In the following code chunk, the first line is identical to the code that produces the scatterplot above except for one addition, the `col` (color) parameter. We mapped the color parameter to the category so the points are colored accordingly.

I've also added a dashed line (`lty`) with a width of 2 (`lwd=2`) to show that the correlation between the likes-to-view and dislikes-to-view ratio.

Finally, I added a legend for our plot to show how the colors of our scatterplot points map to each level of our `category_id` variable.

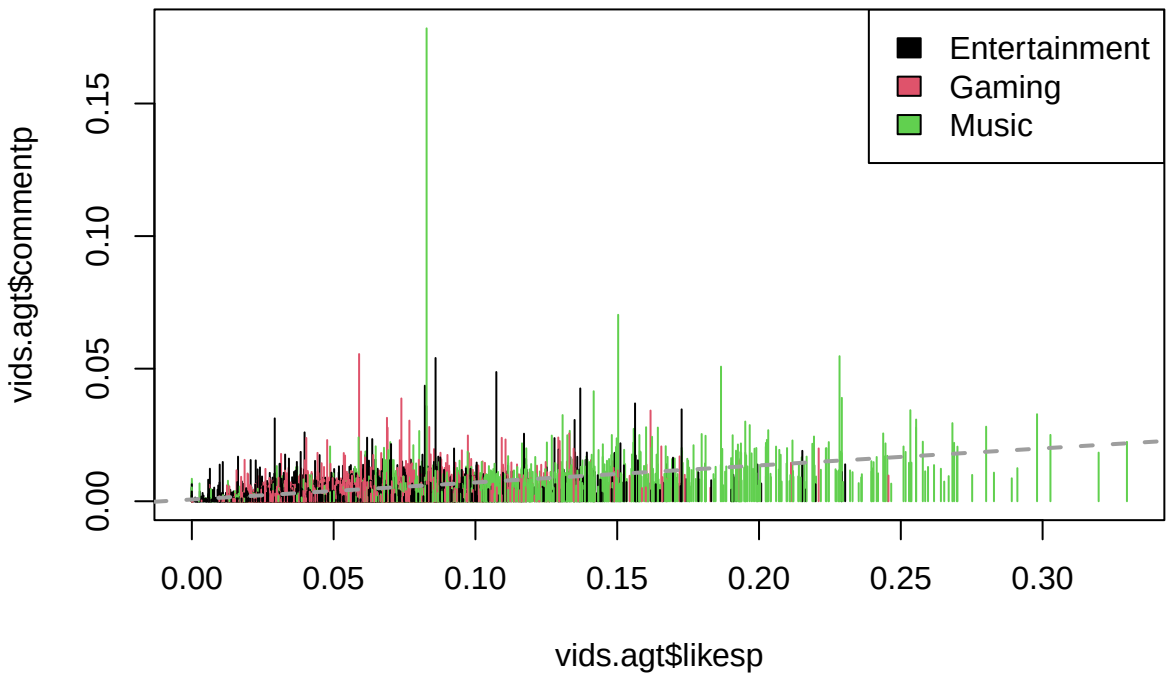
Here's the code:

```
plot(vids.agt$likesp, vids.agt$commentp, col=vids.agt$category_id, pch=19)
abline(lm(vids.agt$commentp ~ vids.agt$likesp), col=8, lwd=2, lty=2)
legend("topright", legend=levels(vids.agt$category_id), fill=1:3)
```



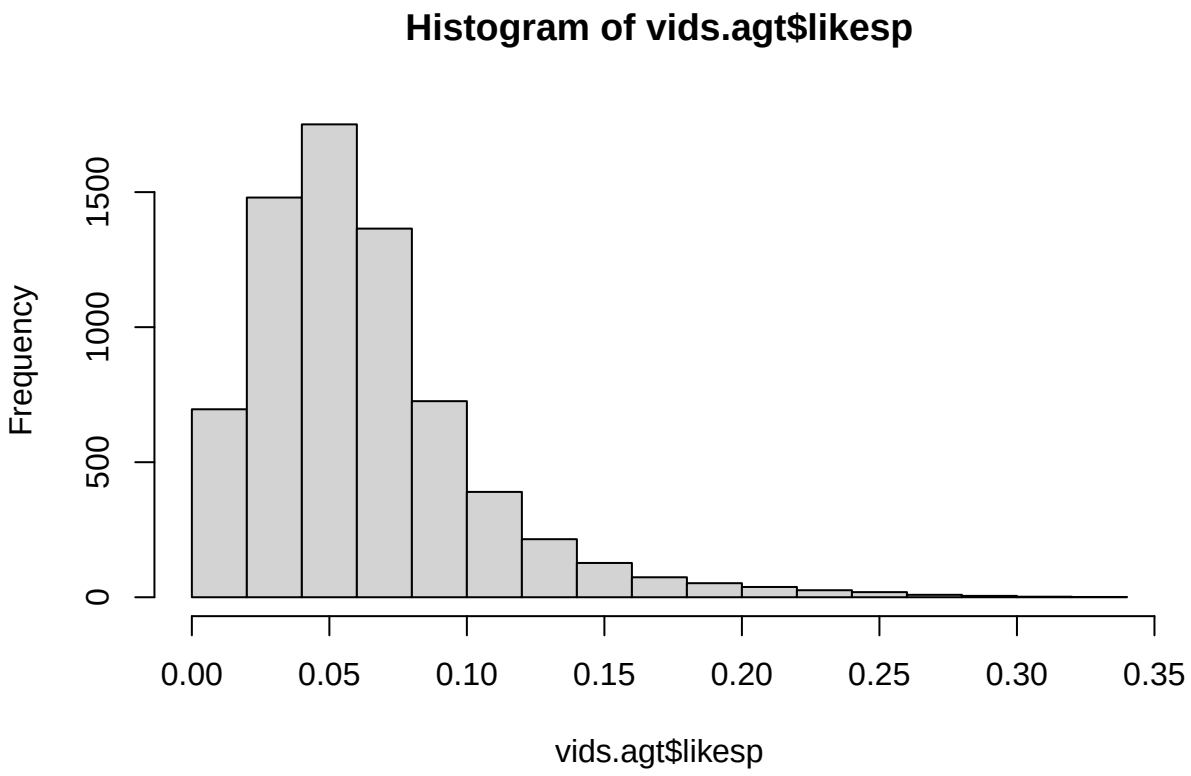
With `plot`, the default for two numerical variables is to plot a scatterplot, but we can override the default parameters with the `type` argument. The following code chunk is identical to the one above, except for the `type="h"` addition:

```
plot(vids.agt$likesp, vids.agt$commentp, col=vids.agt$category_id, type="h")
abline(lm(vids.agt$commentp ~ vids.agt$likesp), col=8, lwd=2, lty=2)
legend("topright", legend=levels(vids.agt$category_id), fill=1:3)
```



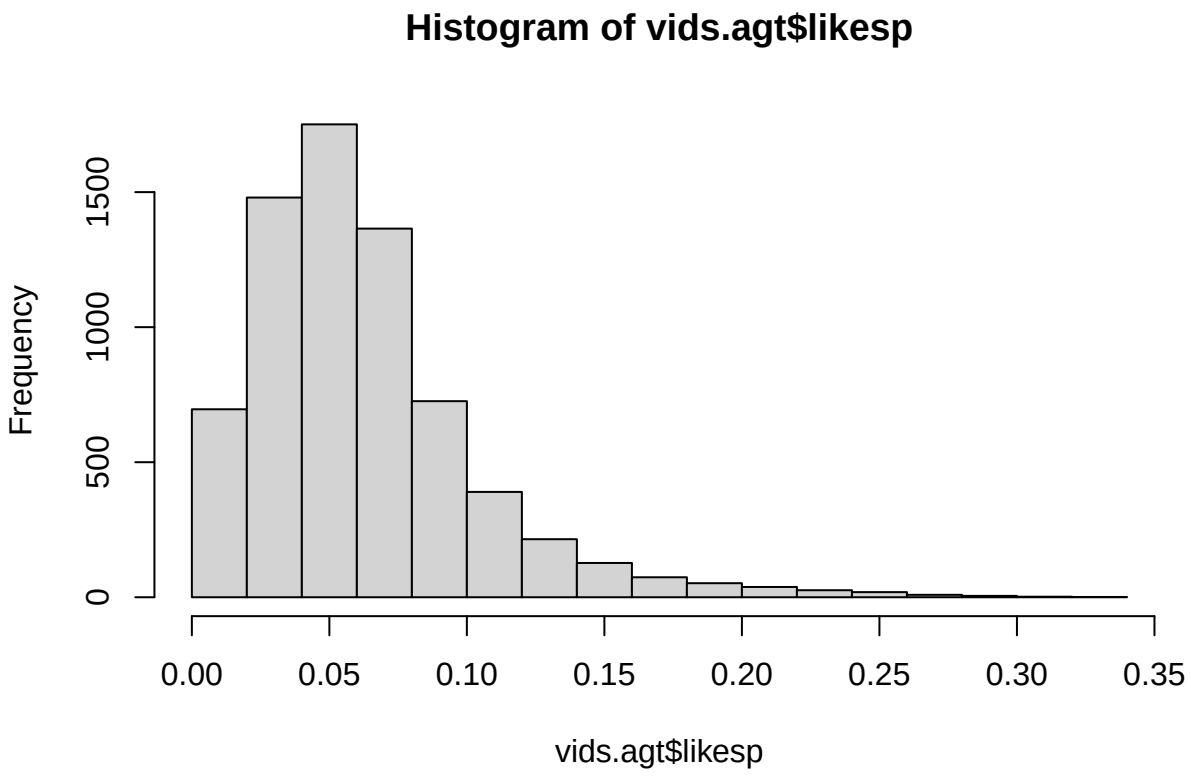
Apart from using `plot()`, we can also create statistical plots using functions such as `hist()`. `hist()` takes a numeric vector and creates a histogram:

```
hist(vids.agt$likesp)
```



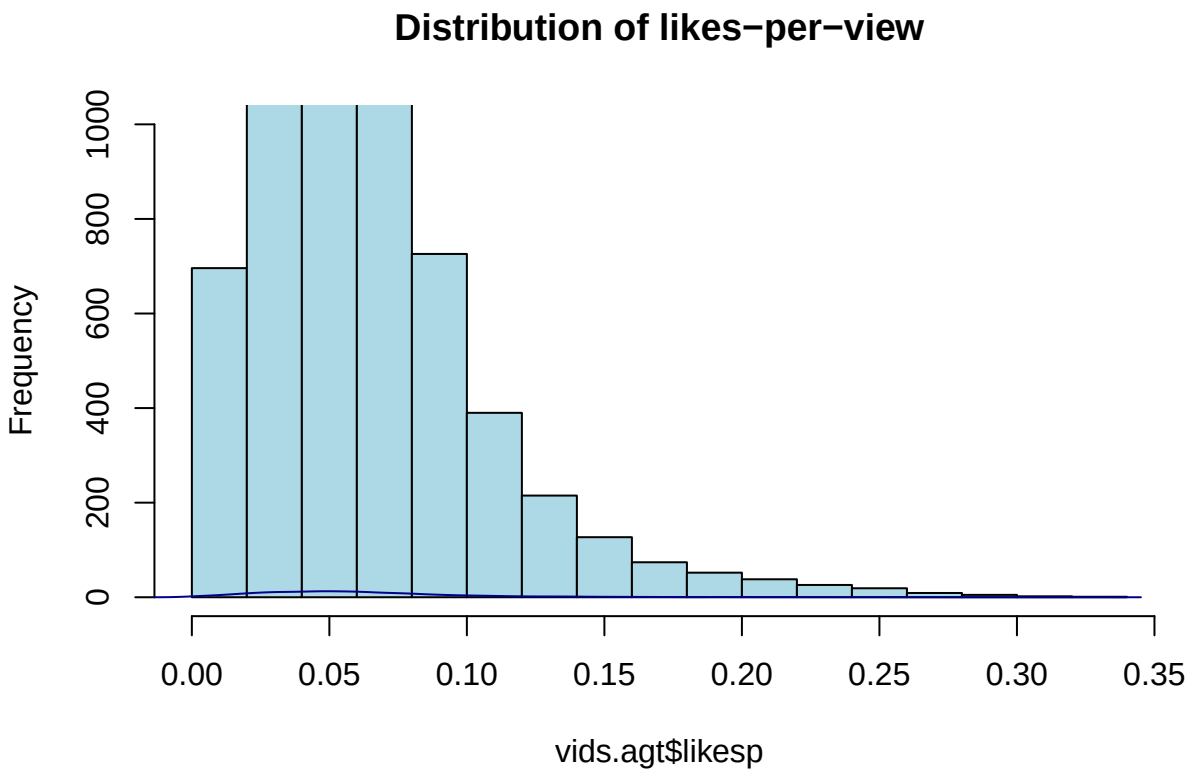
We can additionally use the `breaks` argument to control the number of bins if we were not satisfied with the default values:

```
hist(vids.agt$likesp, breaks=20)
```



Just like how we can add `abline` onto our plot, we can add graphical elements like `lines` onto this histogram too. In fact, let's do that and also use the `main` argument to give our plot a new main title:

```
hist(vids.agt$likesp,
     breaks=20,
     ylim=c(0, 1000),
     col="lightblue",
     main="Distribution of likes-per-view")
lines(density(vids.agt$likesp), col="darkblue")
```



While base plot can be very simple to use, they can be effective too. In fact, with the use of proper coloring, annotation and a little care on the aesthetic touches, you can communicate a lot in a graph using just R's built-in plotting system.

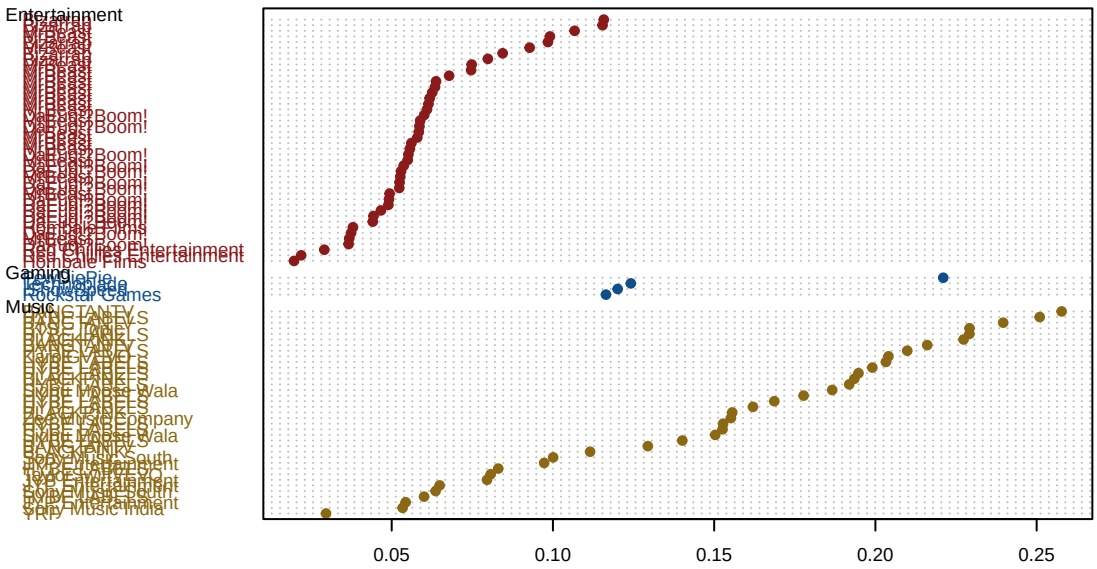
In the following code chunk I'm subsetting from `vids.agt` only trending videos that have more than 1000,000 likes and order it by the likes-to-view variable. I added a new variable, `col` to this new dataframe to be used in my following plot:

```
vids.ags <- vids.agt[vids.agt$likes > 1000000, ]
vids.ags <- vids.ags[order(vids.ags$likesp), ]

# create color specifications for our dotchart
vids.ags$col[vids.ags$category_id == "Music"] <- "goldenrod4"
vids.ags$col[vids.ags$category_id == "Gaming"] <- "dodgerblue4"
vids.ags$col[vids.ags$category_id == "Entertainment"] <- "firebrick4"
```

We're going to create a dot chart (or a Cleveland's Dot Plot) by graphing the likes to view ratio of each trending video in the `vids.ags` dataframe, map `channel_title` to the labels and group these labels by `category_id`.

```
dotchart(vids.ags$likesp, labels=vids.ags$channel_title, cex=.6, pch=19, groups=vids.ags$category_id, c
```



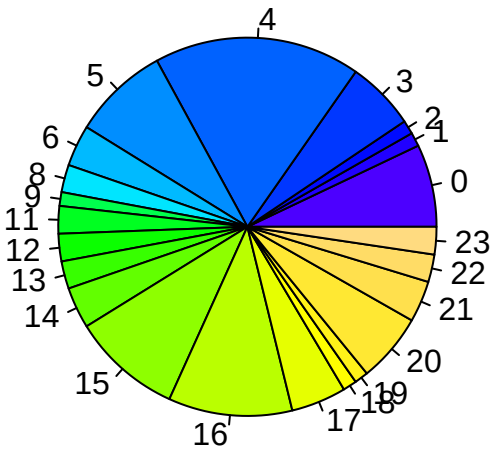
With this we see that between groups, the likes to view proportion (we'll call it "likeability" from now on) of Entertainment, Gaming and Music. Music videos we do see a larger variance in that the top video by likeability is close to 0.30, more than 6 times difference to that of other trending videos in this category. We also observe the rough mean likeability within groups, as well as between them. We would expect Entertainment videos to have ~1 likes per 10 views, and Music videos to have more than that due to the positive skew we observe from above.

Let's talk about another kind of plot, one that most statisticians find cringeworthy for it's undeserved popularity and prevalence in the workplace. Yes, it is the pie chart. In R's official documentation, the pie chart is criticized as being "a very bad way of displaying information [because] the eye is good at judging linear measures and bad at judging relative areas". Almost any data that can be represented in a pie chart can be illustrated with a bar chart or dot chart[2].

If you insist on creating one, here's the code (I've added some colors to make it easier to get a grasp of the measures):

```
pie(table(vids.ags$publish_hour), labels=names(table(vids.ags$publish_hour)), col=topo.colors(24))
```

[2]Full Note on pie charts from the official R Documentation:
"Pie charts are a very bad way of displaying information. The eye is good at judging linear measures and bad at judging relative areas. A bar chart or dot chart is a preferable way of displaying this type of data.



Grammar of Graphics in R

The motivation of ggplot2

`ggplot2` is created by Hadley Wickham in 2005 as an implementation of Lelad Wilkinson's Grammar of Graphics. The idea with Grammar of Graphics is to create a formal model for data visualization, by breaking down graphics into components that could be systematically added or subtracted by the end user.

With `ggplot2`, plots may be created using `qplot()` where arguments and defaults are handled similarly to the base plotting system, or through `ggplot()` where user can add or alter plot components layer-by-layer with a high level of modularity.

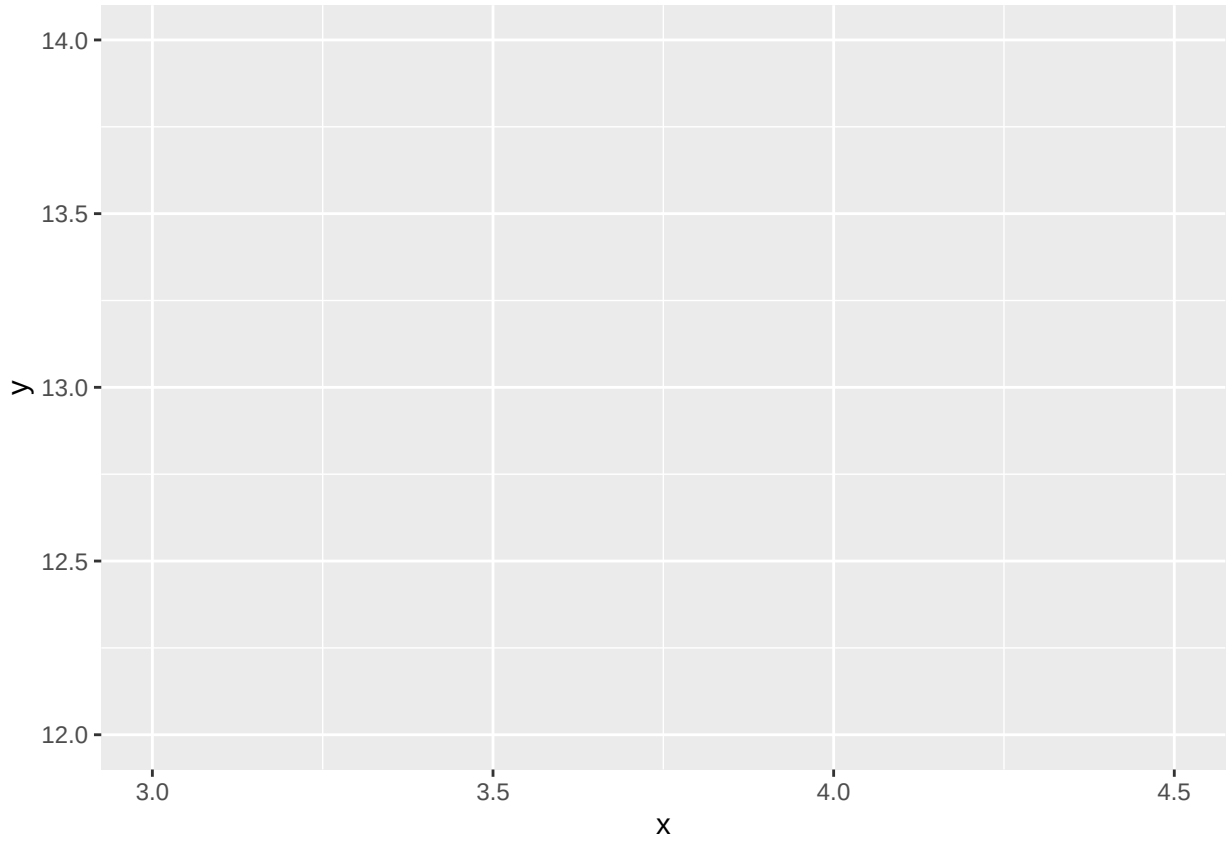
The last point is especially important because it allows the data scientists to work with plots in a system that breaks up these different tasks. Instead of a huge, conceptually flat list of parameters to control every aspect of the plot's final outcome, this system makes plotting a series of distinct task, each focused on one aspect of the plot's final output.

Let us take a look at a simple example, drawing inspiration from the Earthquake incident that happened in the south of Jakarta this week (as of this writing). I've created a dataframe called `gempa`:

```
gempa <- data.frame(
  x=c(3.5,3,4,4.5,4.1),
  y=c(12,14,12.4,12.5,14),
  size=c(14,4,4,6,12)
)
```

And now I'll create a ggplot object using `ggplot()`. Because of the range of my values, this plot will use that and create a plot with these values on each scales (scales, by the way, can be thought of as just the two axis right now). Note that we're just creating a blank plot with no geometry elements (lines, points, etc) on it yet. We save this object as `g`:

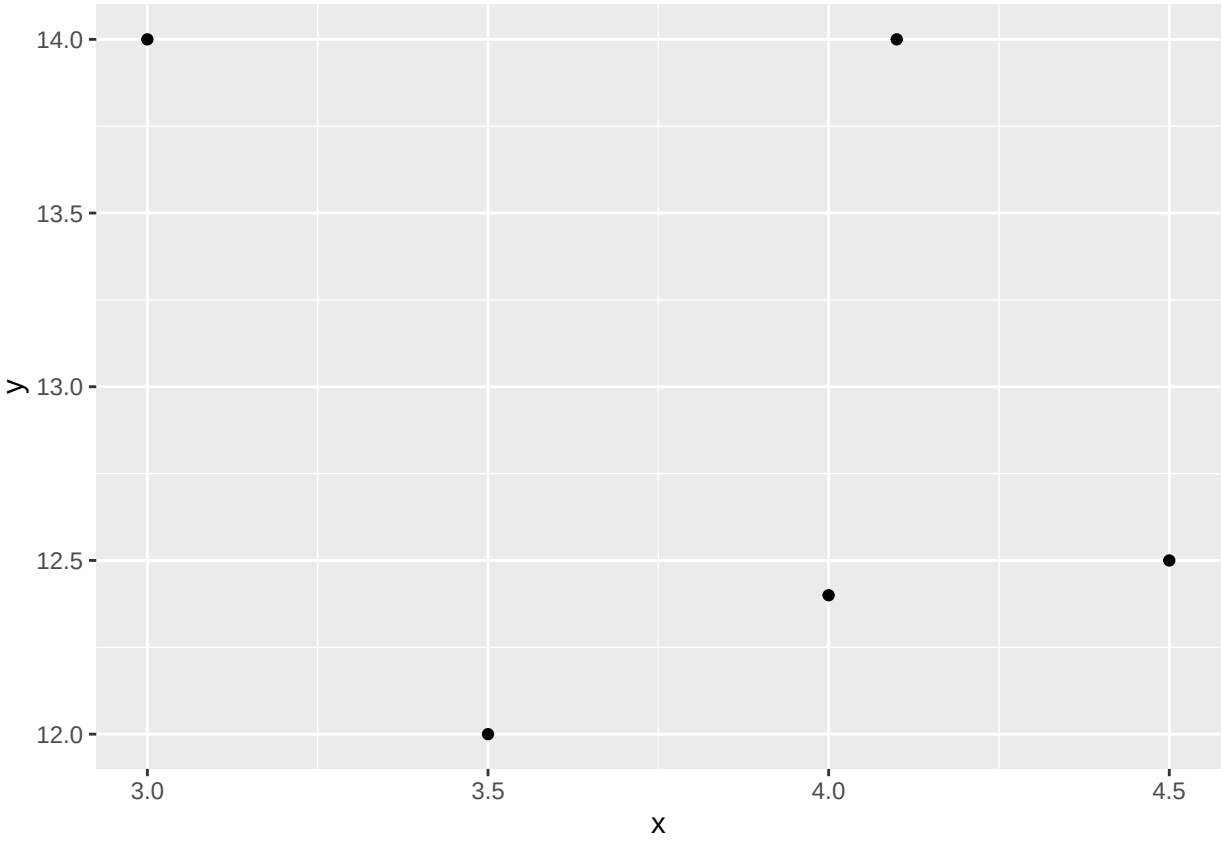
```
g <- ggplot(gempa, aes(x = x, y = y))
g
```



Notice how `ggplot()` takes two arguments: - The data - The `aes` which allow us to specify our mapping of the x and y variables so they are used accordingly by ggplot

Once we created our `ggplot` object (we named it `g`), we can now add a layer onto it using `geom_`. `geom` is ggplot's way of handling geometry, i.e. how the data are represented on the plot. To illustrate this idea, let's add a geom_point and then print the resulting object:

```
g + geom_point()
```



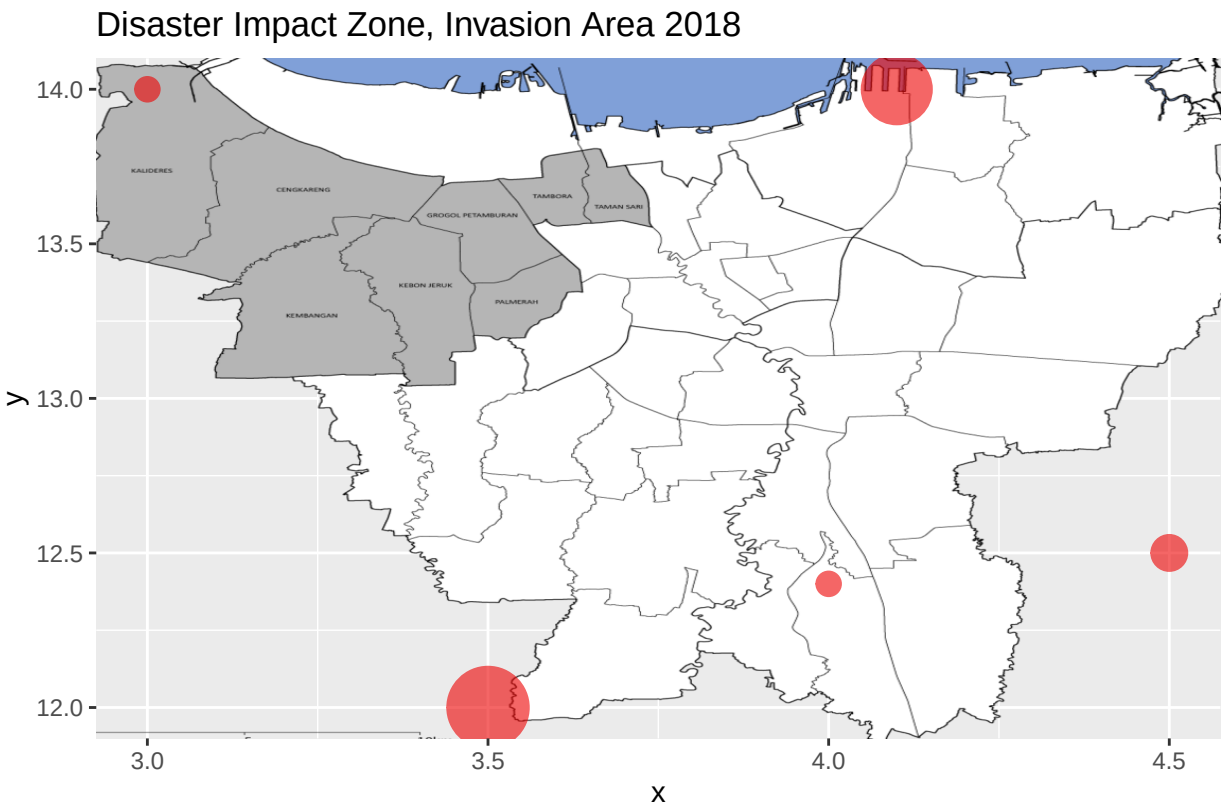
A recap of what we've done so far:

- Creating our ggplot graphics object through `ggplot()`
- We specify 2 arguments in our call to ggplot; It's helpful to note that any argument we pass into `ggplot()` will be used as global options for the plot, i.e. they apply to all layers we add onto that graphics object
- For the second argument we use the `aes()` function, allowing us to map variables from our `gempa` data to aesthetic properties of the plot (in our case we map them to the x and y axis)
- We tell ggplot how we want the graphic objects to be represented by adding (through the "+" operator) our geom layer. Since we added `geom_point`, this is equivalent to adding a layer of scatterplot to represent our x and y variables

As we familiarize ourselves with this system, we will learn to use other functions to obtain a more precise control over the construction of our plot. This could be natively `ggplot` constructs such as scales, legends, geoms and thematic elements or this could be additional constructs that work with `ggplot` through the use of third-party packages. In the following example, we're adding `background_image` to our original plot (`g`) before adding `geom_point` on top of the background image layer. Finally, we add the labels for our title and caption using the `labs` function:

```
library(png)
jak <- png::readPNG('assets/jakarta.png')
g +
  background_image(jak)+
```

```
  geom_point(size=gempa$size, alpha = 0.6, col="red2")+
  labs(title="Disaster Impact Zone, Invasion Area 2018", caption="source: Jakarta Distaster Relief")
```



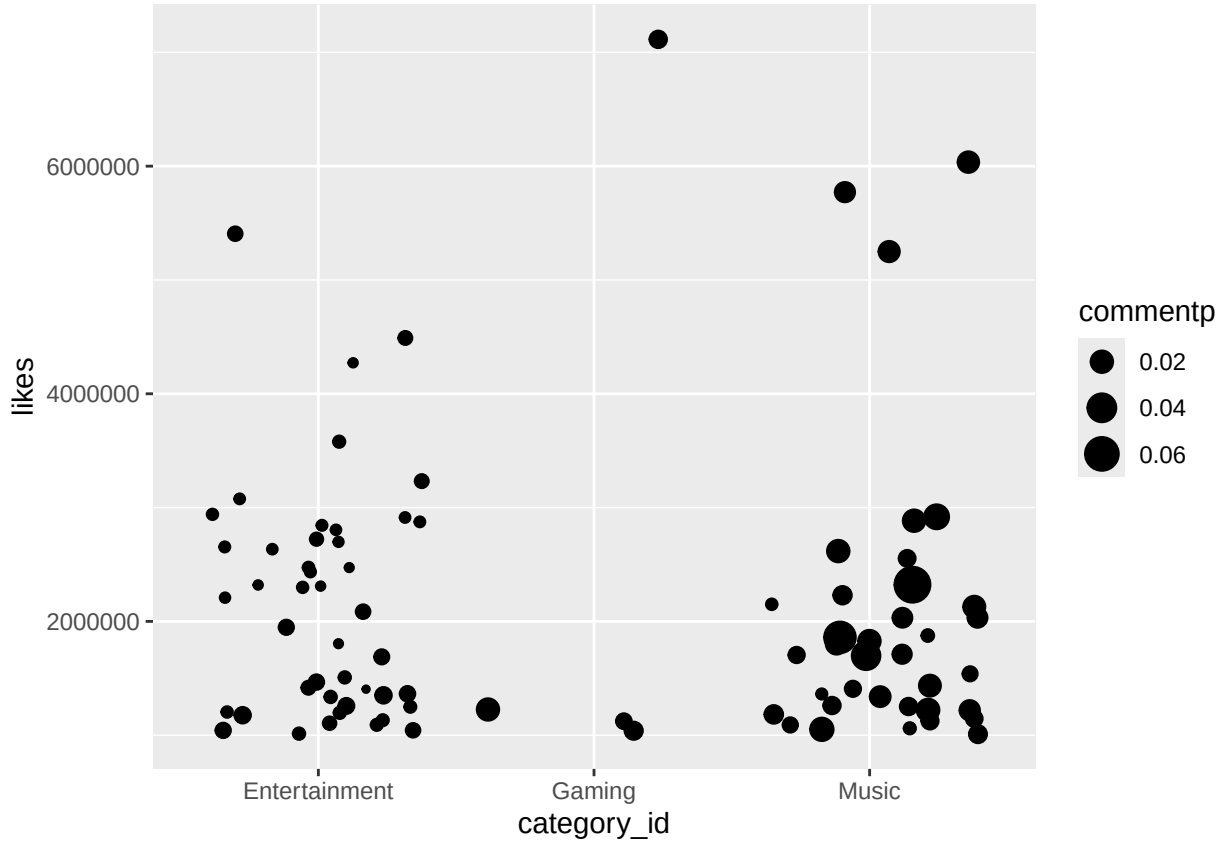
Because of this design philosophy in ggplot, it presents a learning curve that is beginner-friendly and mostly logical. I said beginner-friendly, because as we will see later, all we need to do is to master the starting steps first and not worry about polishing. And with starting steps, this means:
- 1: `ggplot()` with `data` and aesthetics mapping (`aes`) - 2: Add to (1) a single `geom` layer

The State of Trending Videos

Hands-on ggplot: Simple Exploratory Analysis

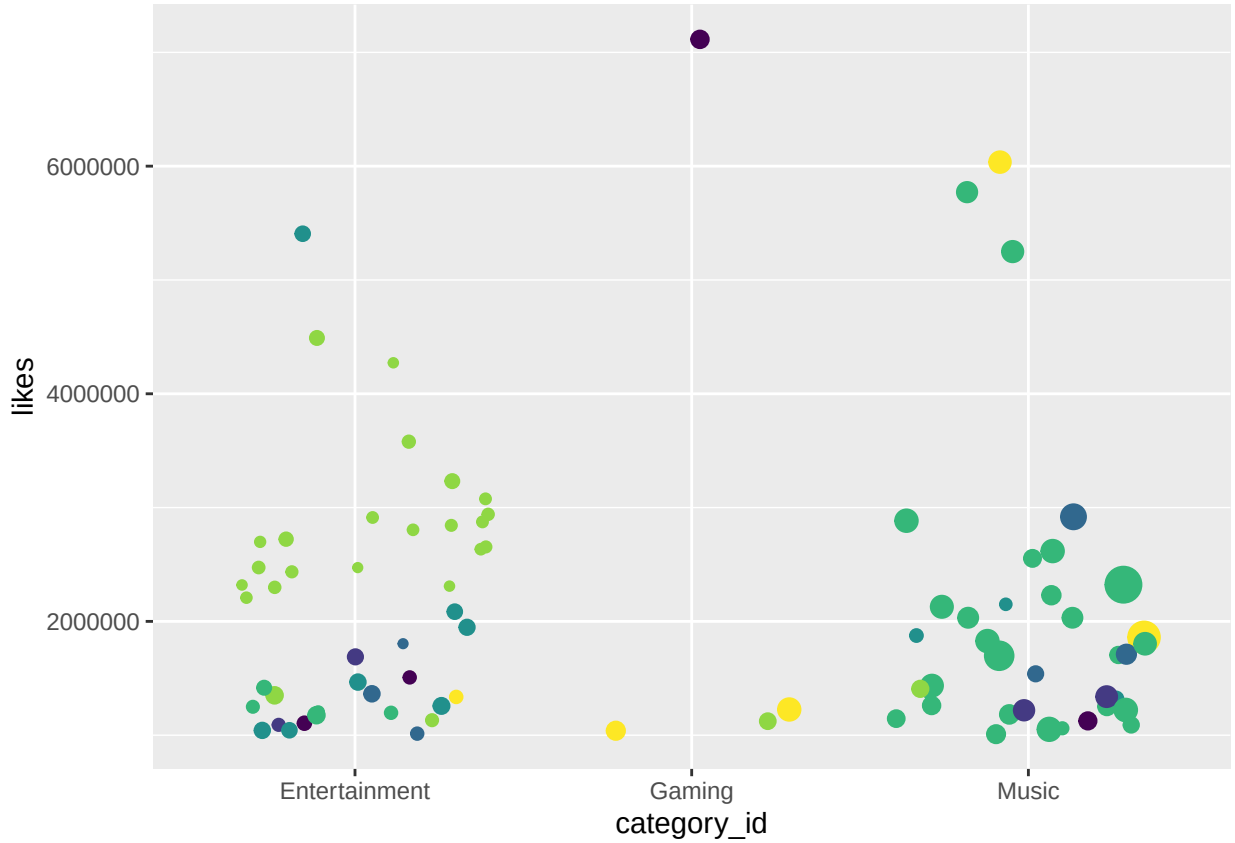
Let's apply what we've learned above to create a simple plot. We will use `vids.ags` as the data, and map our x, y, and size aesthetics to the `category_id`, `likes` and `commentp` respectively.

```
ggplot(data = vids.ags, aes(x=category_id, y=likes, size=commentp))+
  geom_jitter()
```



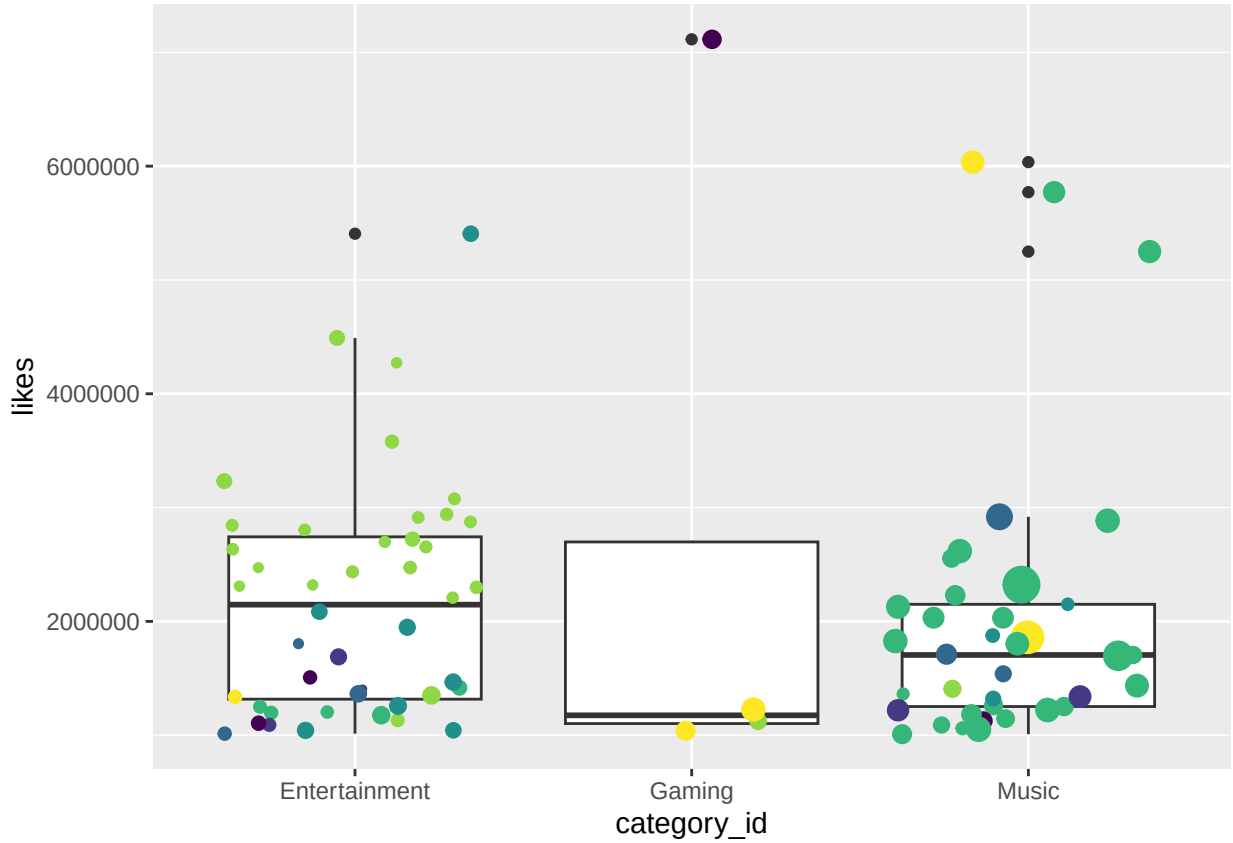
Mostly, I'm copy-and-pasting the code from earlier chunks and then adding ggplot2 elements onto the plot layer-by-layer. Here I mapped the color of my jitter points to the day of week for each of these videos. I also added a `theme` component to specify the position of my legend:

```
ggplot(data = vids.ags, aes(x=category_id, y=likes, size=commentp))+
  geom_jitter(aes(col=publish_wday))+
  theme(legend.position = "none")
```



I hope you're getting the pattern now, but just to drive home the point, let's copy the code above and add one more layer: `geom_boxplot()`. I want the jitter points to be above the boxplot, so I'll add the elements in that order such that the last item will be on the top-most:

```
ggplot(data = vids.ags, aes(x=category_id, y=likes, size=commentp))+
  geom_boxplot()+
  geom_jitter(aes(col=publish_wday))+
  theme(legend.position = "none")
```



Dive Deeper: Give our plot a main title and caption:

Recall that we can use `labs()` to change the labels on our plot and we did just that earlier when we gave our plot a main title and caption. Copy (or re-write) the code above, and gave them an appropriate title and caption. Your code should look something like this:

```
### Copy and Paste or Finish the following pseudo-code

## psuedo-code
## ggplot() +
##   geom_boxplot() +
##   geom_jitter()+
##   theme() +
##   labs(title="Likes to View Comparison", subtitle="Youtube Trending Data, 2018", x="Video Category",

### End of Exercise
```

Hopefully that was a fun introduction to ggplot!

Now let's start thinking about a more useful application of plotting. Imagine you're working with the insights and business intelligence team at a media firm, and were asked to produce a report that illustrates the most prolific producers of trending videos in recent weeks.

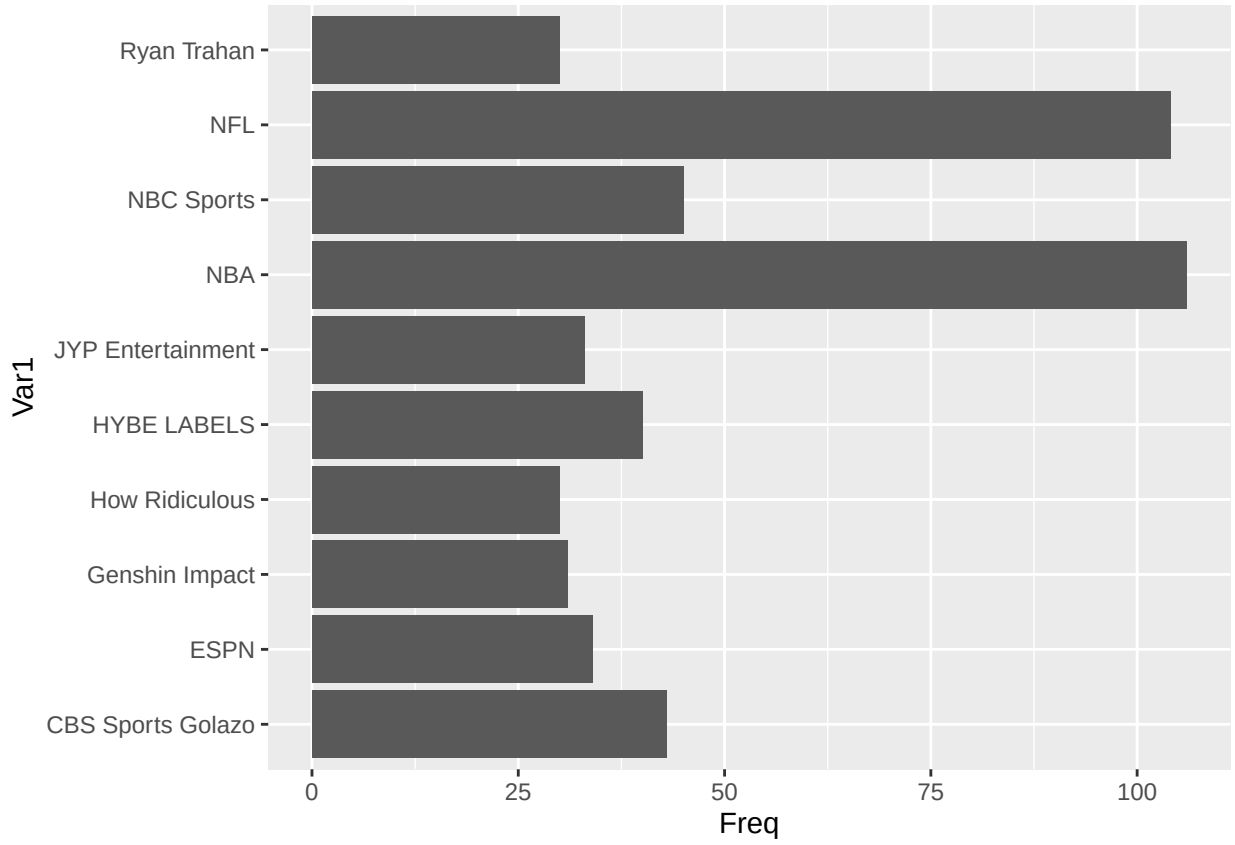
Since we're concerned about the quantity of videos (talking about being *prolific!*) we will create another subset of the full dataframe, but take only the channels that have at least 30 videos being trending!

```
temp1 <- as.data.frame(table(vids.u$channel_title))
temp1 <- temp1[temp1$Freq >= 30,]
temp1 <- temp1[order(temp1$Freq, decreasing = T), ]
head(temp1)
```

```
##                   Var1 Freq
## 2162               NBA  106
## 2220               NFL  104
## 2172        NBC Sports   45
## 498  CBS Sports Golazo   43
## 1342       HYBE LABELS   40
## 954                ESPN   34
```

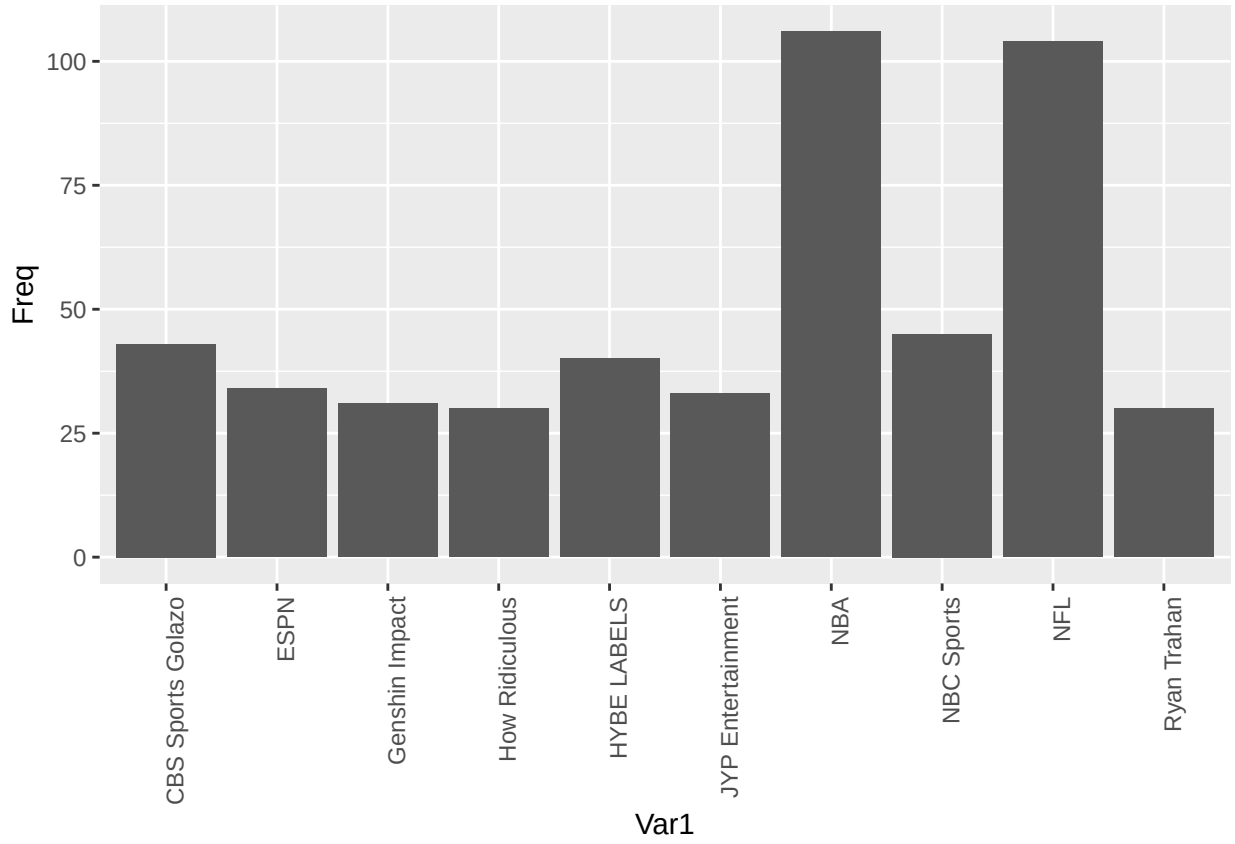
`temp1` is a dataframe with two variables, and contain a list of channels that have at least 10 videos being trending in the observation period. Let's create a column chart (`geom_col`). The category names can be displayed horizontally by placing it on the *x-axis* while the frequency is placed on the *y-axis*. That makes it easier for the user to read and identify the videos that are more prolific than others in producing trending videos:

```
ggplot(temp1, aes(x=Freq, y=Var1))+
  geom_col()
```



If flipping the coordinates (swapping x and y axis) isn't an option, then another approach is to rotate the axis-text for x by 90 degree. This is done with the following code:

```
ggplot(temp1, aes(x=Var1, y=Freq))+
  geom_col()+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



To maximize the time you spent writing code, let's hop into another Dive Deeper challenger. This time we'll create a dataframe with two variables: first is the `category_id` and second being the mean (average) of likes by each of these category. Using what we've learned in Programming for Data Science we create this dataframe using `aggregate()`

Here's the code:

```
temp2 <- aggregate.data.frame(vids.u$likes, by=list(vids.u$category_id), mean)
head(temp2)
```

```
##              Group.1        x
## 1 Autos and Vehicles 33109.61
## 2             Comedy 54942.91
## 3          Education 47312.27
## 4      Entertainment 95430.22
## 5 Film and Animation 77128.52
## 6             Gaming 43989.62
```

Dive Deeper: Can you use `geom_col` to create a column chart that plots the amount of average `comment_count` by category? Refer to earlier exercises if you need to. Add a `labs()` component to label the x- and y-axis appropriately as a bonus.

```
# Write your code here
```

Just to freshen things up, we'll shift the focus from **video categories** to **video producers** for this next exercise. We're now creating a subset of the dataframe that measures the likes-to-views and comment count, but group it on a channel / producer level instead of a categorical level. We'll make sure to take only observations with a non-zero value for the comment count:

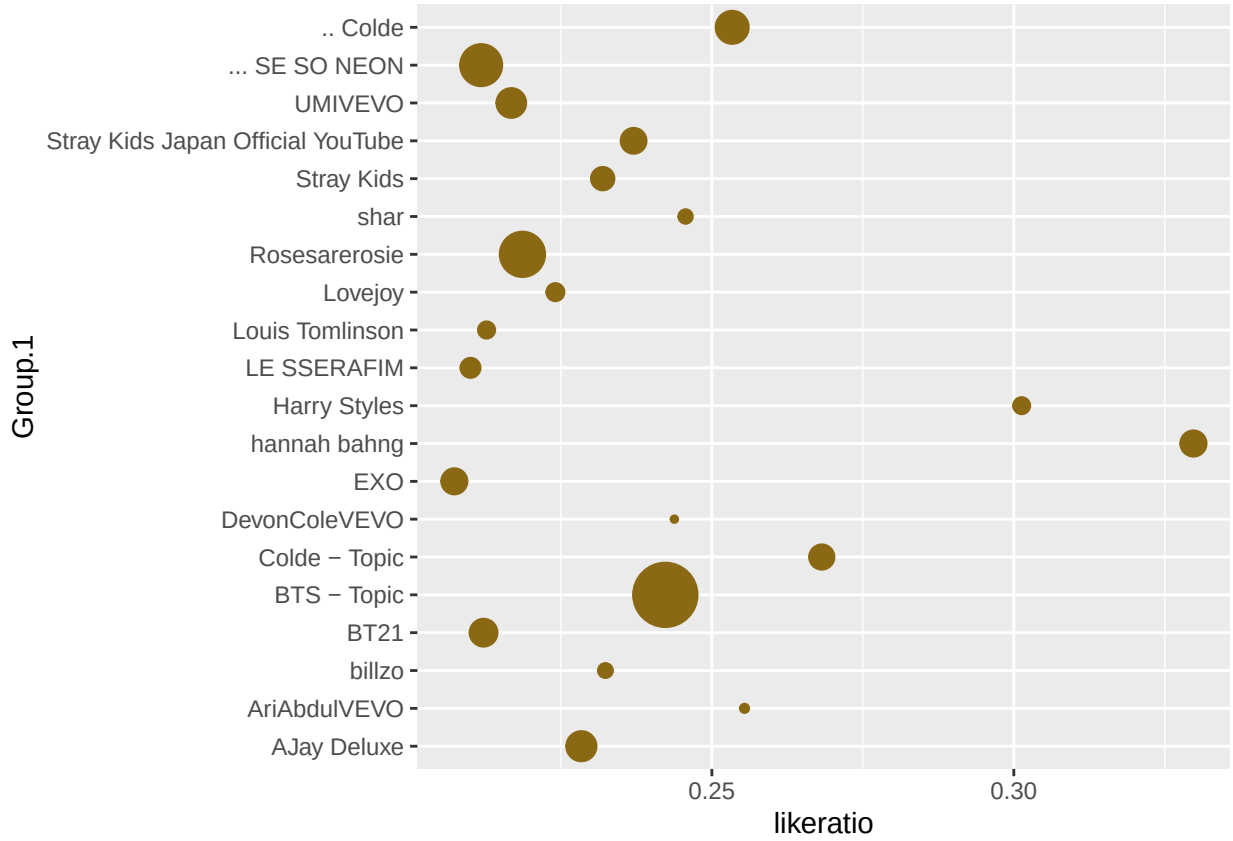
```
temp3 <- aggregate.data.frame(list(likeratio = vids.u$likes/vids.u$views, comment = vids.u$comment_coun
temp3 <- temp3[order(temp3$likeratio, decreasing = T), ]
temp3 <- temp3[temp3$comment != 0, ]
head(temp3)
```

```
##            Group.1 likeratio comment
## 1275  hannah bahng 0.3297434 18874.0
## 1281  Harry Styles 0.3013017  4993.5
## 618  Colde - Topic 0.2681995 17033.0
## 173   AriAbdulVEVO 0.2554054   228.0
## 3536      Colde 0.2533638 35173.0
## 2714          shar 0.2456418  2683.0
```

With the newly created dataframe `temp3`, let's try and create a plot similar to the dot plot we see in an earlier exercise. For now, don't worry about coloring the points or labels and just use a fixed color. However, have the size of each point correspond to the comment count of the videos.

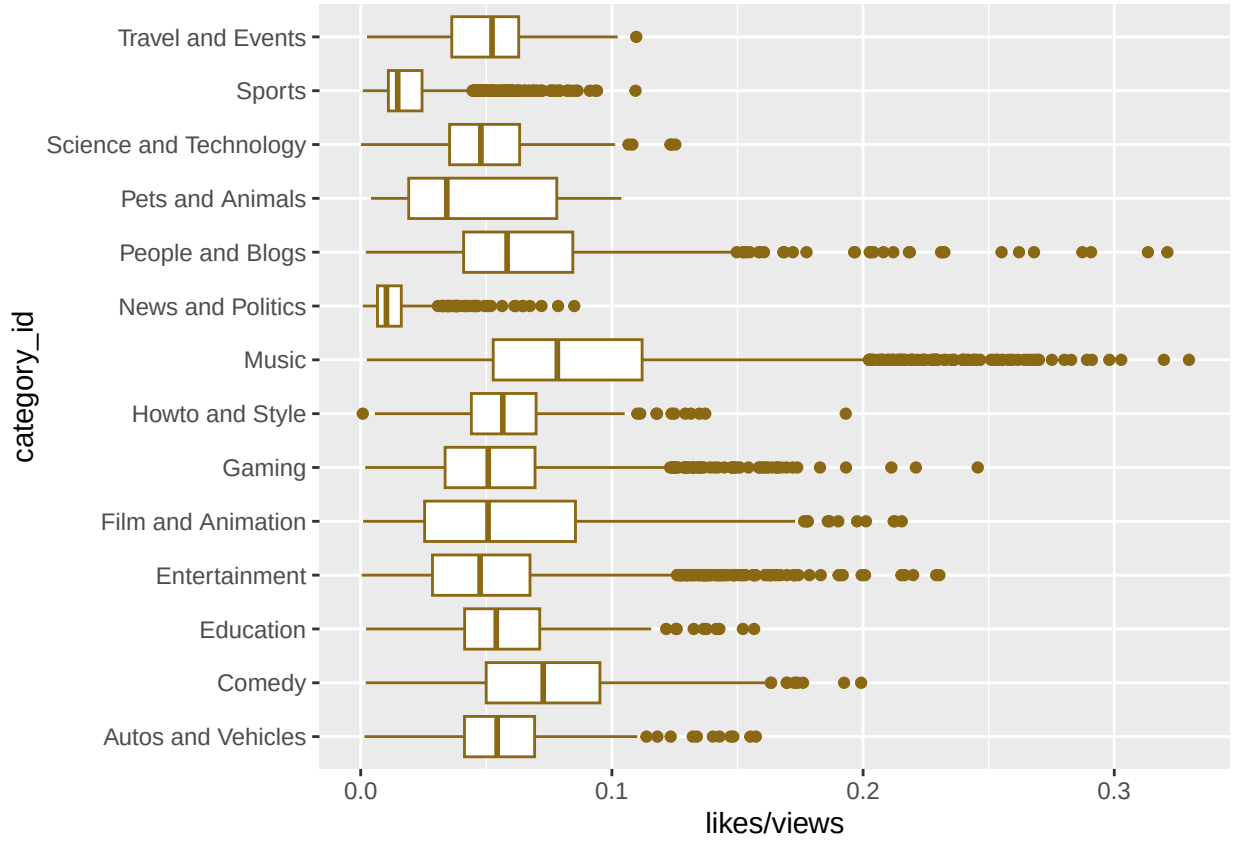
The solution code is here:

```
ggplot(temp3[1:20,], aes(x=likeratio, y=Group.1))+
  geom_point(aes(size=comment), color="goldenrod4",show.legend = F)+
  scale_size(range=c(1,11))
```



We can swap the axis to any other geoms as well to get a better visualization. Such flexbility clearly has much to offer when the data we want to visualize is relatively large, making the right arrangement of the plot's element especially important. Let's see an example below:

```
ggplot(vids.u[vids.u$ratings_disabled == F, ], aes(x=likes/views, y=category_id))+
  geom_boxplot(show.legend = F, color="goldenrod4")
```



A potentially noisy plot, with some clever arrangement, can be made simple and effortless with ggplot. Spend a couple of minutes before hopping into the next section to understand, and practice, what you've learned so far. Hopefully you've seen enough to be convinced of the merits and advantages of this plotting system.

Hands-on ggplot: Multivariate Plots

To take things up a notch, let's see how we can apply what we've learned to create multivariate plots, or plots designed to reveal the relationship among multiple variables, which in turn help us examine the underlying patterns between pairs of variables.

In the following code, I'm reshaping the `vids.u` dataframe from a wide one to a long one, with `category_id` being the ID and measurements being `likes` and and `comment_count`:

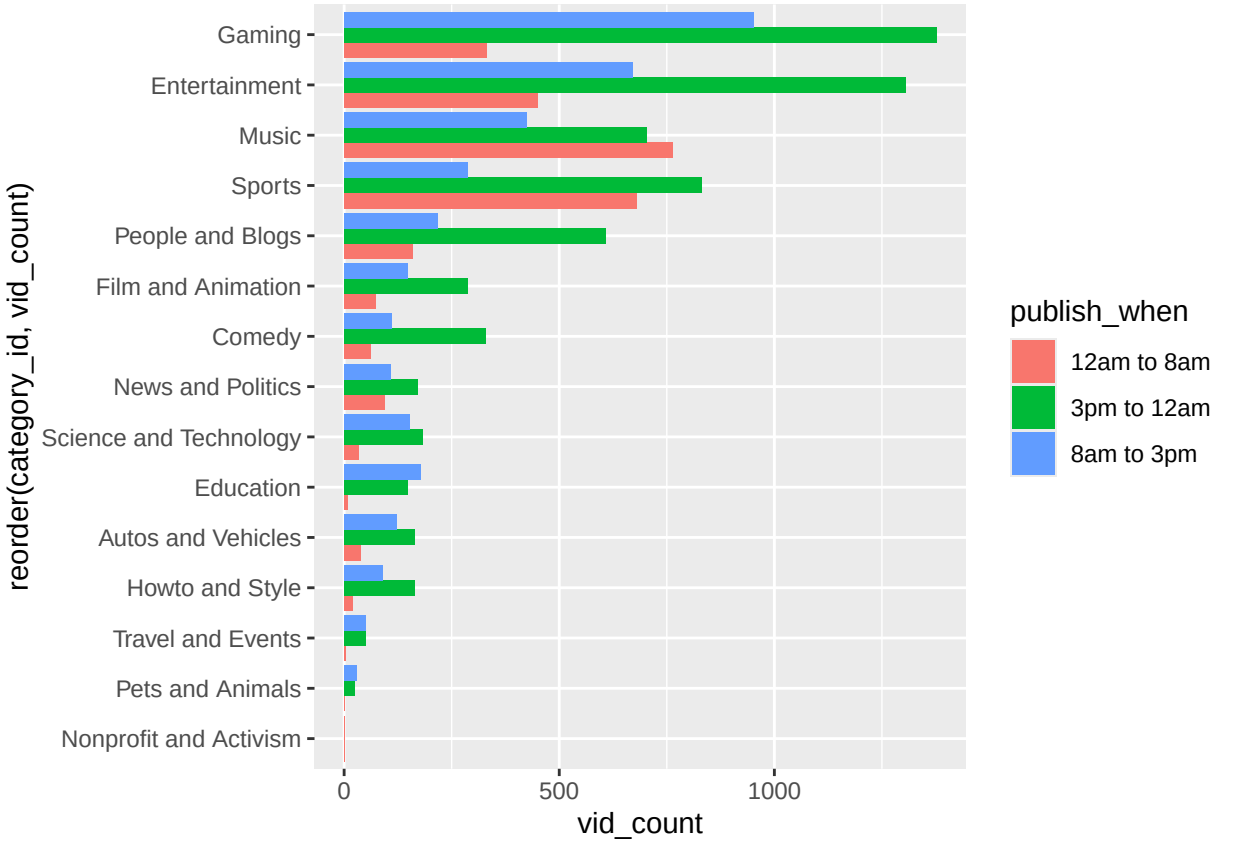
```
# data preparation
vids.select <- aggregate(title ~ category_id + publish_when,
                         data = vids.u,
                         FUN = length)
names(vids.select) <- c("category_id", "publish_when", "vid_count")
head(vids.select)
```

```
##          category_id publish_when vid_count
## 1 Autos and Vehicles  12am to 8am        39
## 2             Comedy  12am to 8am        61
## 3          Education  12am to 8am         7
## 4      Entertainment  12am to 8am       449
```

```
## 5 Film and Animation  12am to 8am        74
## 6              Gaming  12am to 8am       331
```

And using the aggregated dataframe, we can now create `ggplot`. As a start, we'll create a column plot and visualize the `vid_count` of each `publish_when` in each of the `category_id`. We maps the variable `publish_when` to the fill color of the bars, creating grouped bars with different colors based on the `publish_when` variable.

```
ggplot(vids.select, aes(x=vid_count, y=reorder(category_id, vid_count)))+
  geom_col(position="dodge", aes(fill=publish_when))
```



If you have wanted a stacked bar plot instead of side-by-side bar plots, all we need to do is to substitute the `position="dodge"` parameter with `position="stack"`. Try that in the code chunk above and see the resulting plot adjust to that!

Based on the plot above, two categories stand out with the highest number of trending videos: Gaming and Entertainment. This indicates that these categories attract a significant amount of attention and engagement from users on the platform.

Delving deeper into the timing of these trending videos, it is noteworthy that Gaming experiences its peak during the period from 3pm to 12am (midnight). This suggests that gaming-related content, such as gameplay videos, tutorials, and reviews, resonates strongly with viewers during the evening and nighttime hours. It is likely that users, especially avid gamers, actively seek gaming content as a form of entertainment and relaxation during this time frame.

On the other hand, the category with the lowest number of trending videos is pets and animals. This implies that content featuring pets, animal-related events, or educational videos about animals may have

comparatively lower viewership and engagement on the platform. This could be due to the nature of the content and the specific interests of the platform's users.

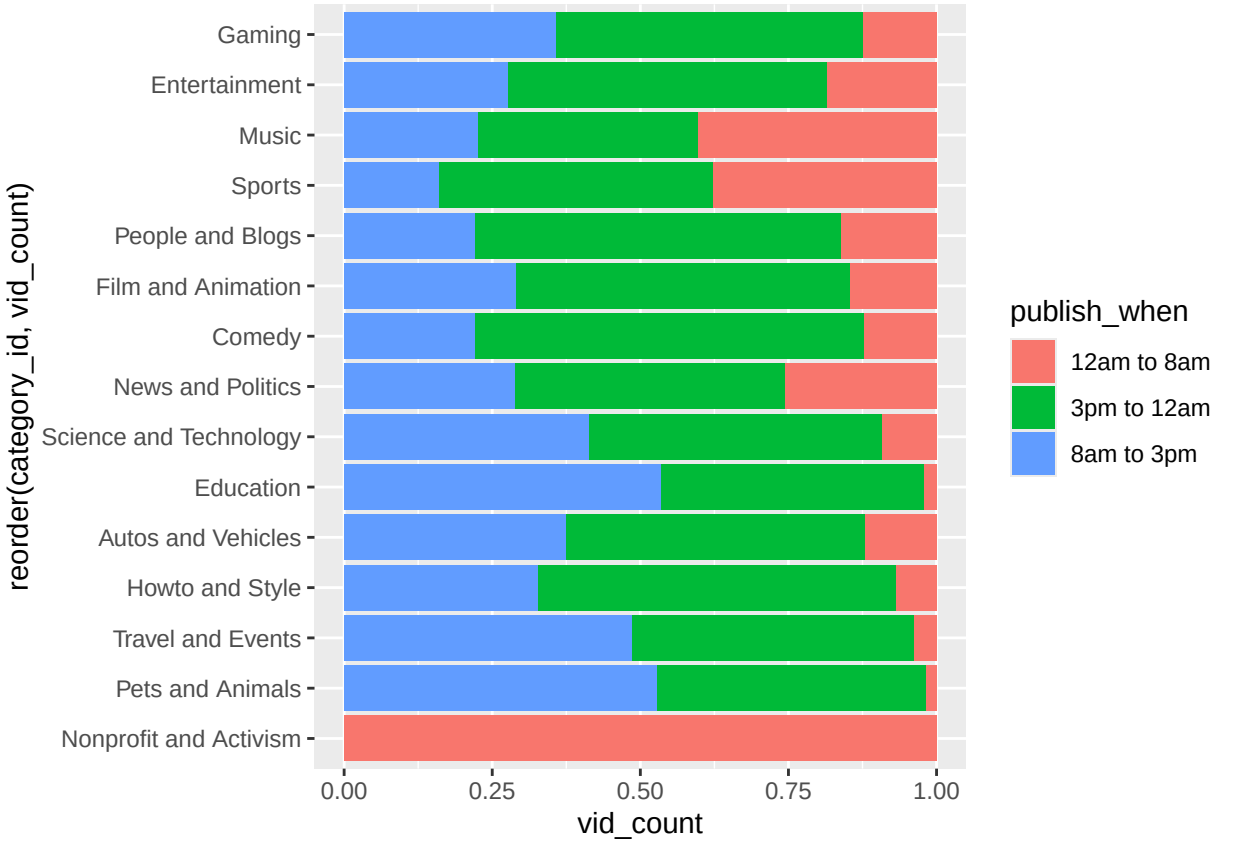
These insights shed light on the popularity of Gaming and Entertainment content, with Gaming thriving during the evening and nighttime hours. Additionally, the relatively low number of trending videos in the pets and animals category indicates a lesser degree of attention and engagement for such content. Content creators and marketers can leverage these insights to better understand user preferences and tailor their strategies accordingly, focusing on the thriving Gaming and Entertainment categories while considering ways to enhance engagement in the pets and animals category.

Next, we will get another insight by using the `position = "fill"` argument to visualize the data with proportions. We can gain insights into the relative distribution of video categories based on their proportions within the dataset.

This visualization method allows us to observe the proportion of each video category relative to the total, showcasing the composition of the dataset in a stacked column format.

The insights we can obtain from this visualization include identifying the categories that contribute the most and the least to the overall dataset. It enables us to compare the relative popularity or prevalence of different video categories based on their proportions.

```
ggplot(vids.select, aes(x=vid_count, y=reorder(category_id, vid_count)))+
  geom_col(position="fill", aes(fill=publish_when))
```



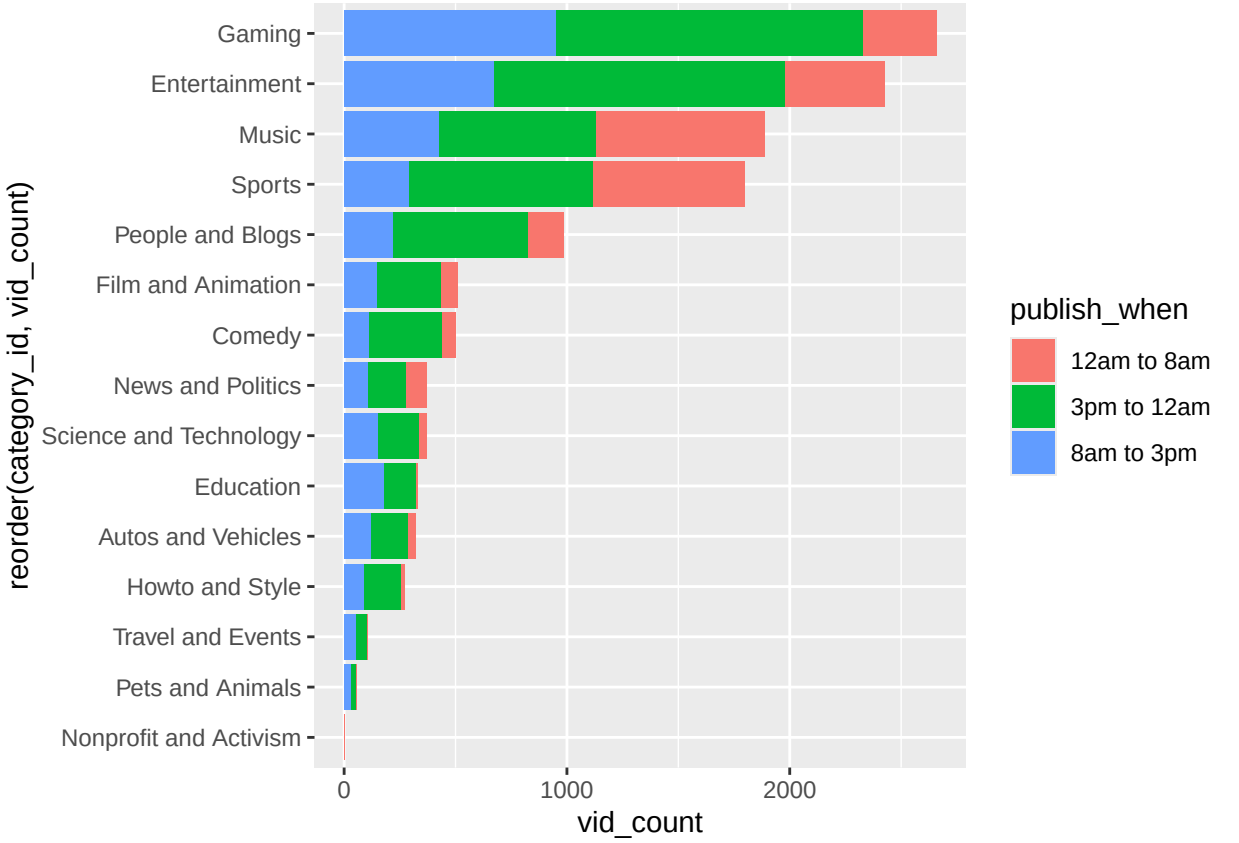
Additionally, we can observe any shifts or changes in the distribution of video categories over time or across different variables, as the stacked columns represent the proportional composition within each category.

By using `position = "stack"`, we can easily observe the total value of each category and identify which category has the highest or lowest overall count. This is particularly valuable when we are interested in analyzing the absolute values or comparing the relative sizes of the categories.

This method is suitable when the focus is on the total quantity rather than the proportion or percentage of each category. It allows for a clear visual representation of the distribution of values across categories and enables easy identification of categories that have larger or smaller totals.

By creating a stacked column chart with `position = "stack"`, we can compare the overall view counts of different categories on YouTube. This visualization would enable us to identify which categories have the highest or lowest total views, giving insights into the popularity and engagement levels of each category.

```
ggplot(vids.select, aes(x=vid_count, y=reorder(category_id, vid_count)))+
  geom_col(position="stack", aes(fill=publish_when))
```



For instance, if we find that the Entertainment category has the highest total view count, while the Pets and Animals category has the lowest, it suggests that viewers are more inclined to watch and engage with entertainment-related content on YouTube compared to videos featuring pets and animals.

This visualization approach can be valuable when analyzing the popularity or viewership of video categories on YouTube, providing a clear visual representation of the total view counts for each category and allowing for easy comparisons to identify trends or patterns in viewer preferences.

Overall, by using `position = "stack"` in the context of a YouTube dataset, we can visualize and compare the total view counts across different video categories, helping us gain insights into the relative popularity and engagement levels of each category on the platform.

Hands-on ggplot: Multiple groups of data

To add a bit of diversity in our exercise, we'll now subset from our original data any videos produced by the channel **NewJeans** and we will also compute the mean on each of these videos across the different days it was featured on - these are represented as `newjeans` and `newjeans.agg` respectively:

```r
newjeans <- vids[vids$channel_title == "NewJeans", ]
newjeans.agg <- aggregate.data.frame(newjeans$views, by = list(newjeans$title), mean)
names(newjeans.agg) <- c("title", "mean")
newjeans.agg
```

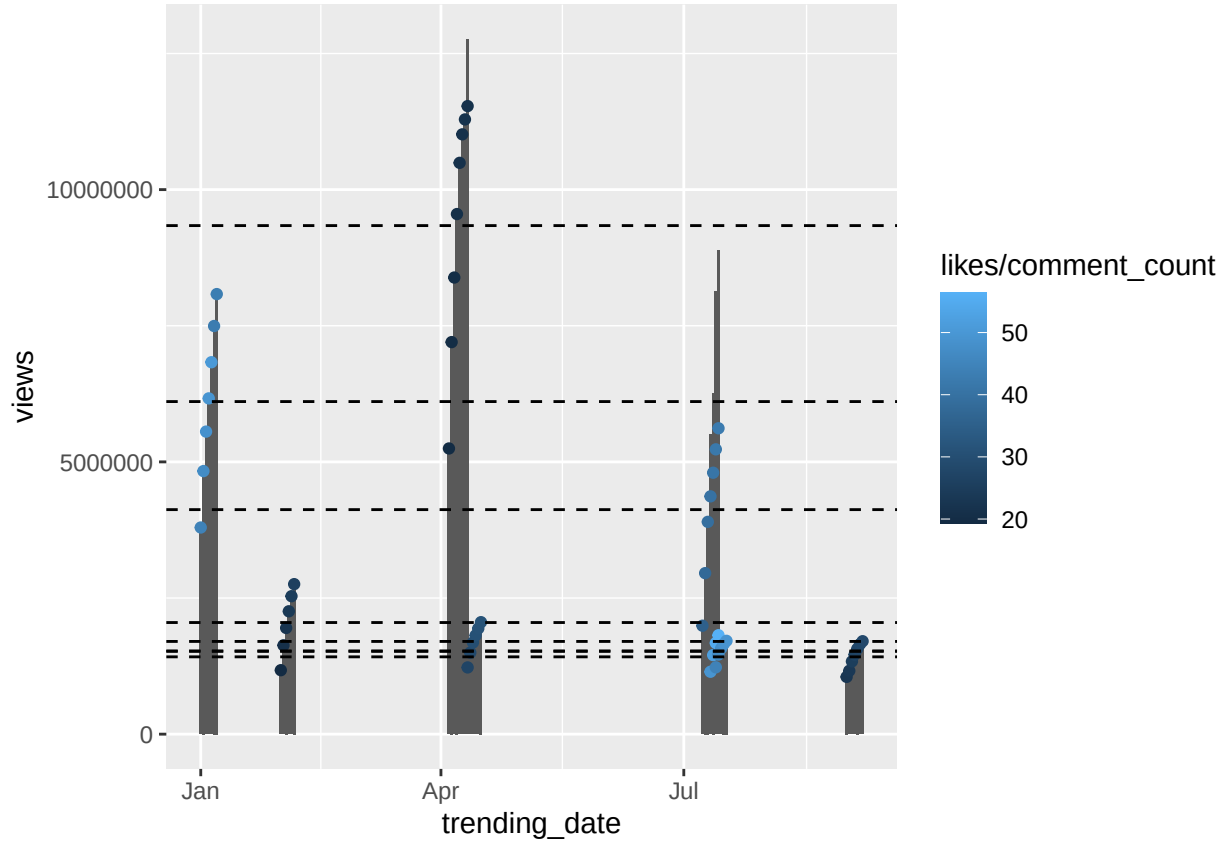
```
##                                                                 title
## 1                [By Jeans] 'V - Rainy Days' Cover by DANIELLE | NewJeans
## 2 [Making Jeans] NewJeans (  ) 'Ditto' & 'OMG' Recording Behind(Final)
## 3                              NewJeans (  ) 'Ditto' Performance Video
## 4                             NewJeans (  ) 'New Jeans' Dance Practice
## 5                             NewJeans (  ) 'Super Shy' Dance Practice
## 6                 NewJeans (  ) 'Super Shy' Dance Practice (Fix ver.)
## 7                                   NewJeans (  ) 'Zero' Official MV
## 8                               NewJeans (  ) 'Zero' Performance Video
##      mean
## 1 1419620
## 2 2050278
## 3 6108129
## 4 1527339
## 5 4123381
## 6 1522296
## 7 9339407
## 8 1703048
```

Sometimes we want to visualize relationships between variables in multiple subsets of the data - in these situations a particularly elegant solution is to have the plots appearing in panels defined in our plot construction process. These types of plots are called facet plots and is an incredible feature of ggplot. Let's see how we can take advantage of that!

Here we're creating a plot using the `hybe` dataset, first by creating the columns and then adding a second layer of points, and finally horizontal lines `geom_hline` (an equivalent for vertical line would be `geom_vline`) for each mean in our `hybe.agg` dataframe. By passing in `data=hybe.agg`, it overrides the initial data (`hybe`) allowing us to add plot elements using aesthetic from a different dataframe. Because we have 5 rows of means, we would expect 5 dashed lines (`linetype=2`).

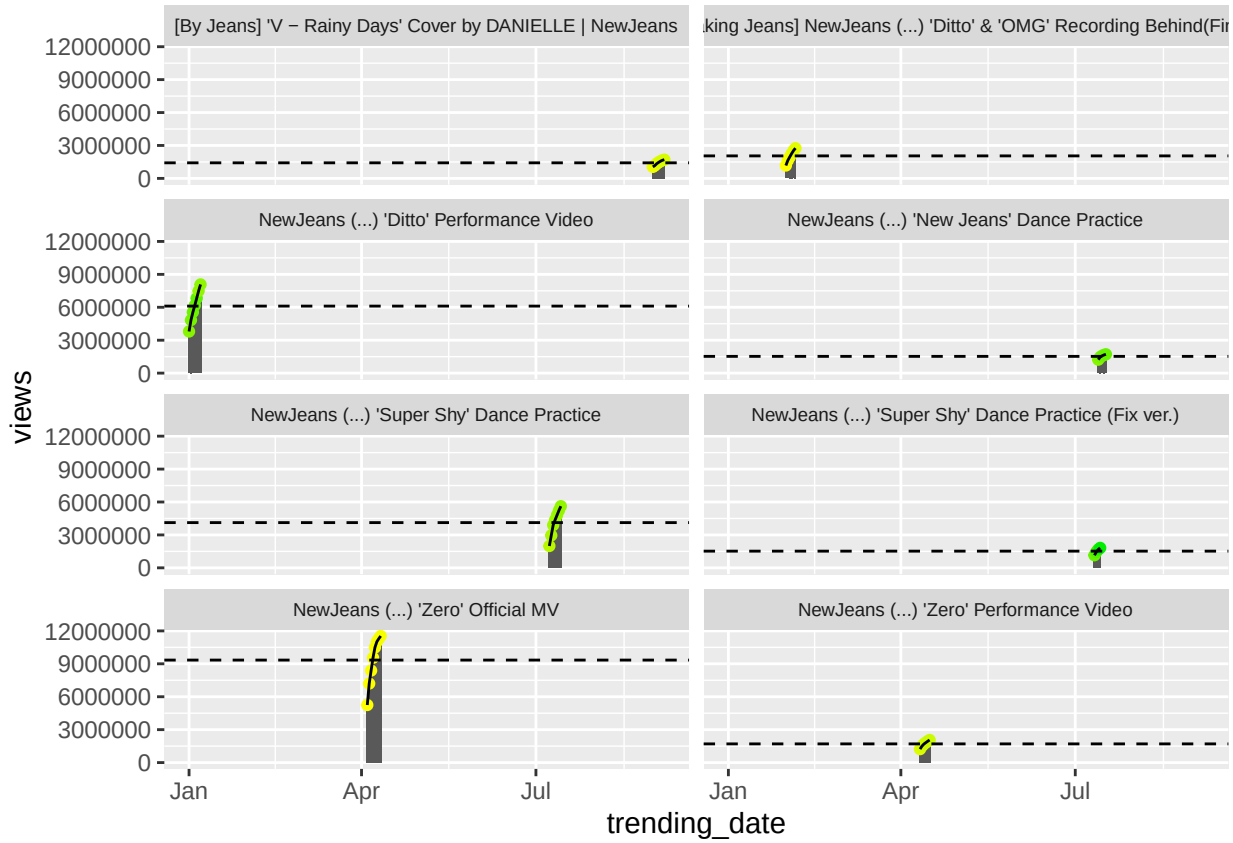
```r
g1 <- ggplot(newjeans, aes(x=trending_date, y=views))+
  geom_col()+
  geom_point(aes(col=likes/comment_count))+
  geom_hline(data=newjeans.agg, aes(yintercept=mean),linetype=2)

g1
```



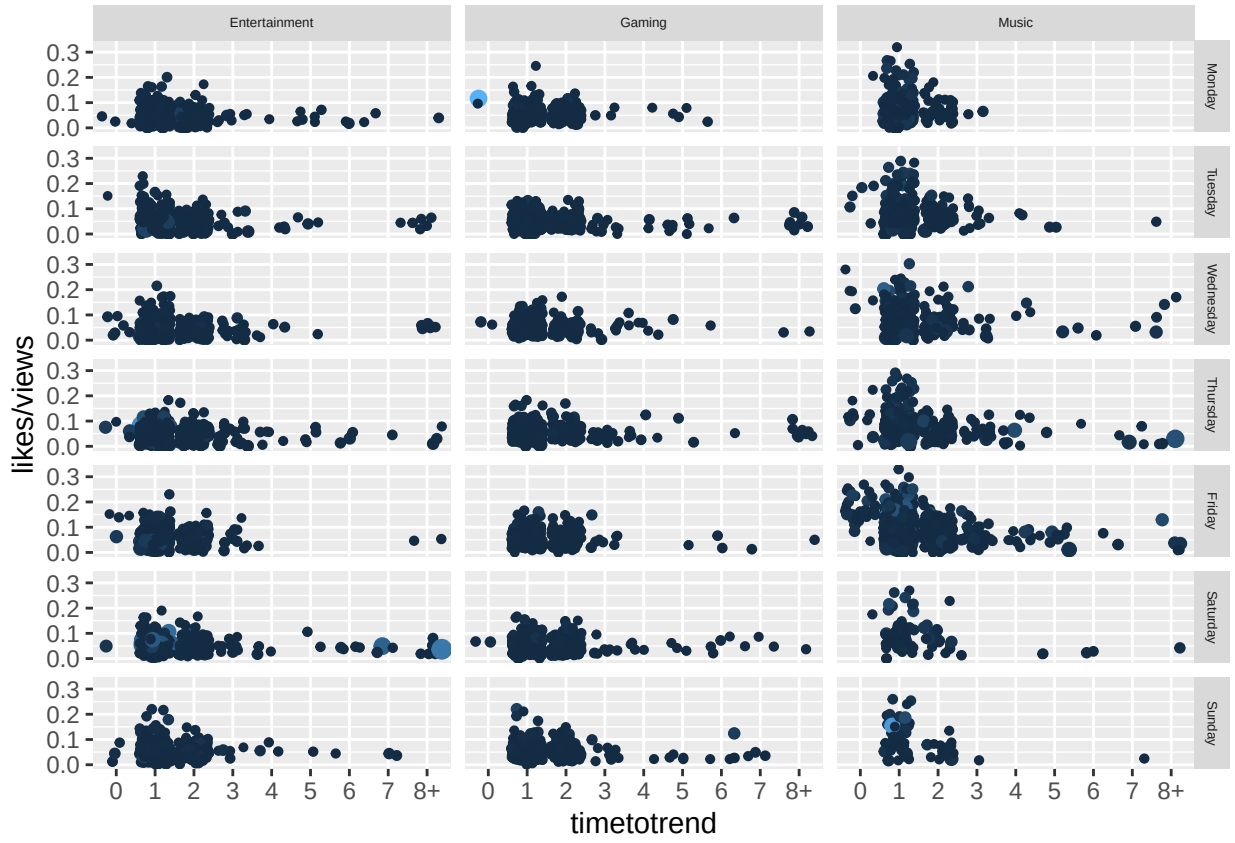
It isn't a bad start, but it may have been a better design decision to break the plot up into panels by each video so we can have a better observation of each video's trend. This could be established using `facet_wrap()`. `facet_wrap` wraps our plots into a multirow panel of plots. Here we'll specify the number of columns to be 2 using `ncol=2`:

```
g1 +
  geom_line(col="black")+
  facet_wrap(~as.factor(title), ncol=2)+
  scale_color_gradient(low="yellow", high="green2")+
  theme(legend.position = "none",
        strip.text=element_text(size=7))
```



In a similar way to `facet_wrap`, we can use `facet_grid` to specify which variables are used to split our plots along the rows and columns. If we want the row to be the day of the week and columns to represent each category, we could achieve that by adding `facet_grid(publish_wday~as.factor(category_id))`:

```
ggplot(data=vids.agt, aes(x=timetotrend, y=likes/views))+
  geom_jitter(aes(size=views, col=likes))+
  facet_grid(publish_wday~as.factor(category_id))+
  scale_size(range=c(1,3))+
  theme(legend.position = "none",
        strip.text=element_text(size=5))
```



Dive Deeper: Using facet_grids or facet_wraps on Disney Plus Imagine working with Disney Plus and tasked to produce a visualization using data of their trending videos in past recent weeks. Use either `facet_wrap` or `facet_grid` in your visualization.

```
vids.disney <- vids.u[vids.u$channel_title == "Disney Plus", ]
vids.disney$sentiment <- vids.disney$likes/vids.disney$comment_count
vids.disney$title <- as.factor(vids.disney$title)
head(vids.disney)
```

```
##       trending_date
## 45598    2023-08-19
## 51270    2023-09-17
## 51792    2023-09-20
## 59453    2023-10-28
## 63391    2023-11-17
##                                                                  title
## 45598 We've Been Expecting You | Percy Jackson and the Olympians | Disney+
## 51270                    Goosebumps | Official Trailer | Disney+ and Hulu
## 51792                  Percy Jackson and The Olympians | Teaser | Disney+
## 59453    Doctor Who 60th Anniversary Specials | Official Trailer | Disney+
## 63391         Percy Jackson and The Olympians | Official Trailer | Disney+
##       channel_title      category_id        publish_time   views likes dislikes
## 45598   Disney Plus People and Blogs 2023-08-18 16:12:12  279134 17757        0
## 51270   Disney Plus People and Blogs 2023-09-14 15:00:14  268325  9236        0
## 51792   Disney Plus People and Blogs 2023-09-19 16:00:09 1617637 71537        0
## 59453   Disney Plus People and Blogs 2023-10-25 18:30:03  379656 10265        0
```

```
## 63391   Disney Plus People and Blogs 2023-11-16 14:00:33 2166665 64701        0
##       comment_count comments_disabled ratings_disabled video_error_or_removed
## 45598          1775             FALSE            FALSE                  FALSE
## 51270           976             FALSE            FALSE                  FALSE
## 51792          5173             FALSE            FALSE                  FALSE
## 59453          1392             FALSE            FALSE                  FALSE
## 63391          4118             FALSE            FALSE                  FALSE
##       publish_hour publish_when publish_wday timetotrend sentiment
## 45598           16  3pm to 12am       Friday           1 10.003944
## 51270           15   8am to 3pm     Thursday           3  9.463115
## 51792           16  3pm to 12am      Tuesday           1 13.828919
## 59453           18  3pm to 12am    Wednesday           3  7.374282
## 63391           14   8am to 3pm     Thursday           1 15.711753
```

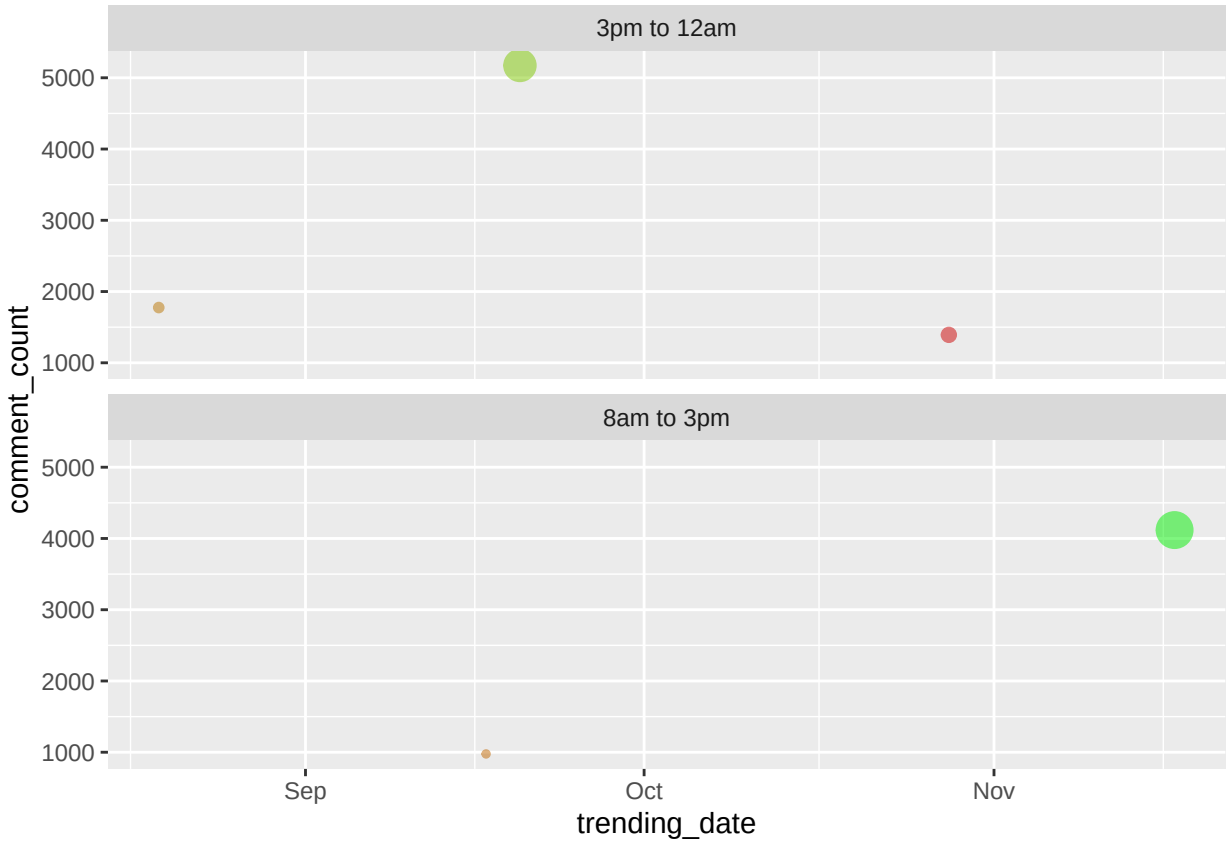
```
### Write your answer here and execute the code to produce your visualization

# -- Scroll up to refer to earlier code and copy / paste any if you like
# -- Try and attempt this challenge by discussion with your neighbor if necessary but do not blindly co
```

While I'll show you a reference answer here below, I strongly recommend you only use this as a reference upon completion of the **Dive Deeper** exercise.

```
disneyp <- ggplot(vids.disney, aes(x=trending_date, y=comment_count, color=sentiment))+
  # geom_point is a reference, geom_col or other geoms (sensible choices) would work too
  geom_point(aes(size=views),show.legend = F, alpha=0.5)+
  scale_color_gradient(low="red3", high="green2")+
  # facet_wrap by publish when is a reference, any other variables (sensible choices) would work
  facet_wrap(~publish_when, ncol=1)

disneyp
```



There are some pretty interesting points in the above plot. The bright green one near "December 01" is an interesting video we may want to further study, and the one that crosses 12,500 comment count is just about as interesting. One quick fix is to replace each points with `geom_text` or use `geom_label` to label them:

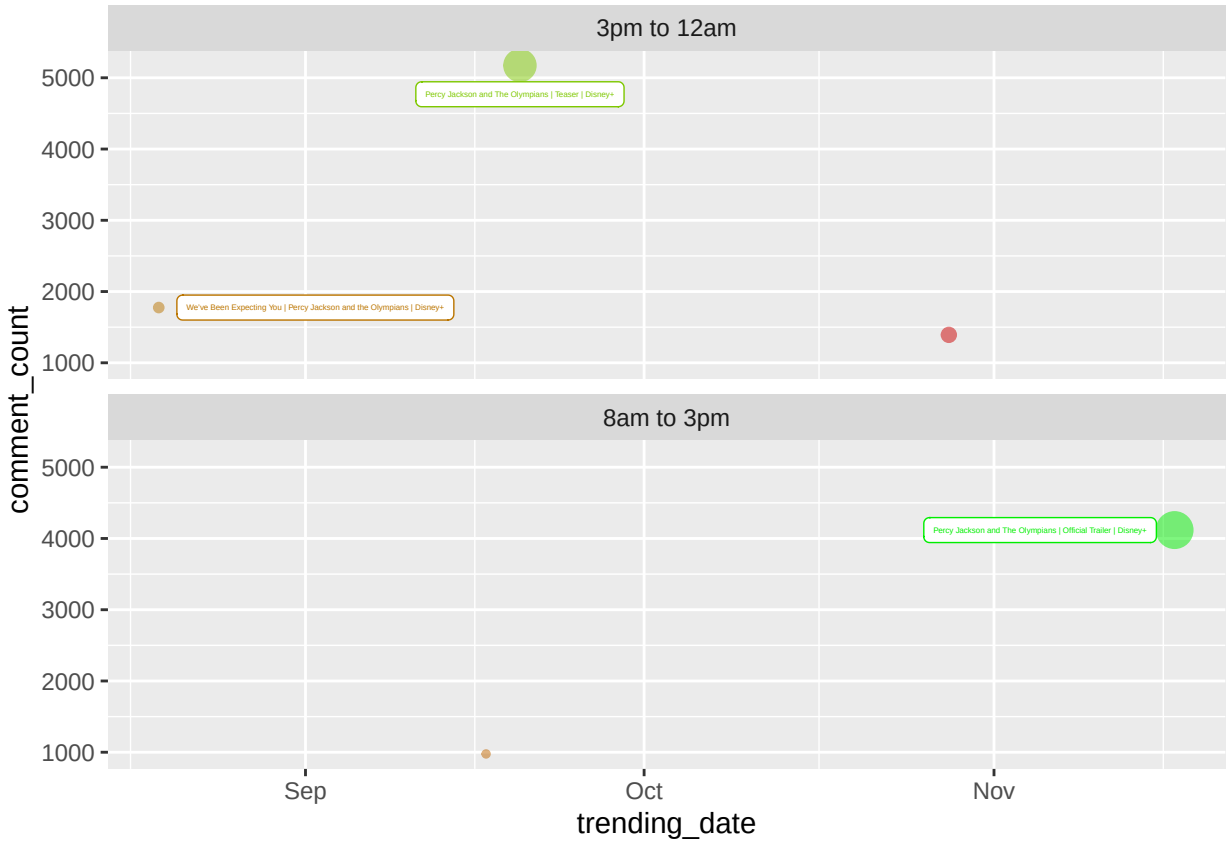
```
disneyp +
  geom_label(aes(label=title), size=1.4, nudge_y = 800) +
  theme(legend.position = "none")
```

Notice that a lot of these labels overlapped with each other and are generally just noisy to look at (a phenomenon sometimes called "overplotting"). We have a few ways to fix that. First, is to only plot points that could be potentially interesting. Second, is to have the text or labels "repel" each other, creating more space between these different points.

This can be achieved using a library called `ggrepel`:

```
library(ggrepel)
disney.pop <- vids.disney[vids.disney$views >= 400000 | vids.disney$sentiment >= median(vids.disney$sen

disneyp +
  geom_label_repel(aes(label=title), size=1.2, data=vids.disney[vids.disney$title %in% disney.pop,], bo
  theme(legend.position = "none")
```



Now this result is a lot neater, and definitely more comfortable on the eyes. We see that "Teaser | Percy Jackson and the Olympians | Disney+" is the most engange video.

Can you discuss any other findings or propose any preliminary observations from the plot above? How would you improve the plot above?

Learn-by-building Module: Data Visualization

To help us get ready to the learn-by-building assignment, we'll walk through a simple exercise together. There isn't any new concept being introduced, but rather a start-to-finish recap of the data visualization process.

First, we'll have to subset any videos within the Educations category and take only the ones that have more than 10 trending videos during the explanatory period. We run the following code and see that 6 channels / publishers satisfy that condition:

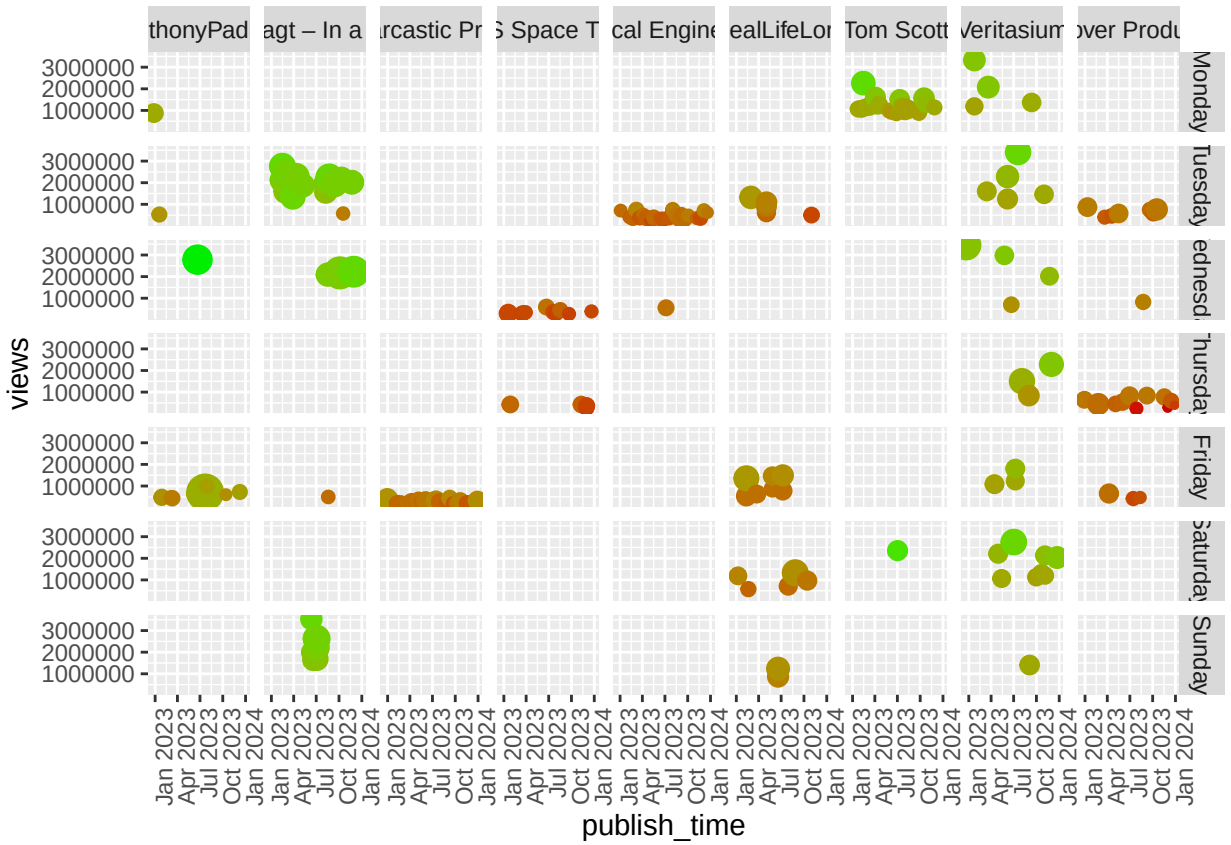
```
education <- vids.u[vids.u$category_id == "Education", ]
education <- aggregate(trending_date ~ channel_title, education, length)
education <- education[education$trending_date > 10, ]
education <- education[order(education$trending_date, decreasing = T),]
education
```

```
##                  channel_title trending_date
## 65                   Veritasium           28
## 35   Kurzgesagt - In a Nutshell           26
## 50        Practical Engineering           24
```

```
## 66          Wendover Productions            23
## 64                     Tom Scott            21
## 51                   RealLifeLore           19
## 46 Overly Sarcastic Productions             18
## 47                 PBS Space Time           17
## 6                  AnthonyPadilla           11
```

We can now use `vids.u$channel_title %in% news$channel_title` as our data source, indicating that we only wish to create our ggplot using channels that are in the list of 6 news channels above. The rest of the code is relatively straightforward:

```
ggplot(data=vids.u[vids.u$channel_title %in% education$channel_title,], aes(x=publish_time, y=views))+
  geom_point(aes(col=log(likes), size=comment_count))+
  facet_grid(publish_wday~channel_title)+
  scale_color_gradient(low="red3", high="green2")+
  theme(legend.position = "none",
        axis.text.x = element_text(angle = 90, hjust = 1))
```



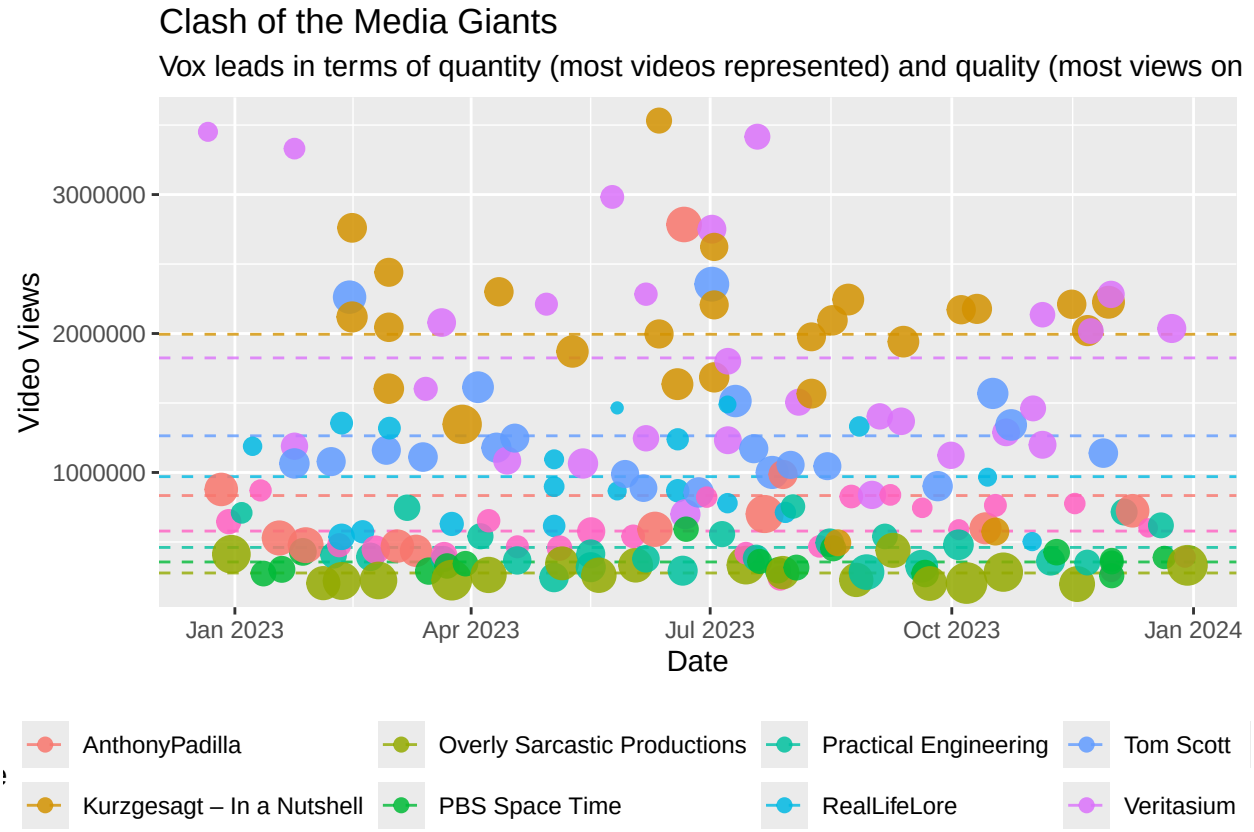
We don't always need to use a facet plot, if the information you wish to convey doesn't require that. The following plot does not tells us the sentiment or comment count of each video, but does enough to tell us about each channel's representation in the trending videos.

```
education.agg <- vids.u[vids.u$channel_title %in% education$channel_title, ]
education.agg <- aggregate(views ~ channel_title, education.agg, mean)
names(education.agg) <- c("channel_title", "mean_views")
education.agg
```

```
##                  channel_title mean_views
## 1               AnthonyPadilla   834097.7
## 2   Kurzgesagt - In a Nutshell  1994708.8
## 3 Overly Sarcastic Productions   276780.9
## 4               PBS Space Time   356088.0
## 5        Practical Engineering   460615.8
## 6                RealLifeLore   970101.9
## 7                   Tom Scott  1263845.5
## 8                  Veritasium  1824435.1
## 9         Wendover Productions   578519.2
```

The plot:

```
ggplot(data=vids.u[vids.u$channel_title %in% education$channel_title,], aes(x=publish_time, y=views))+
  geom_hline(data=education.agg, aes(yintercept=mean_views, col=channel_title), linetype=2, alpha=0.8)+
  geom_point(aes(size=likes/views, col=channel_title), stroke=1, alpha=0.85)+
  guides(size=F)+
  labs(title="Clash of the Media Giants", x="Date", y="Video Views", subtitle="Vox leads in terms of qua
  theme(legend.position = "bottom")
```

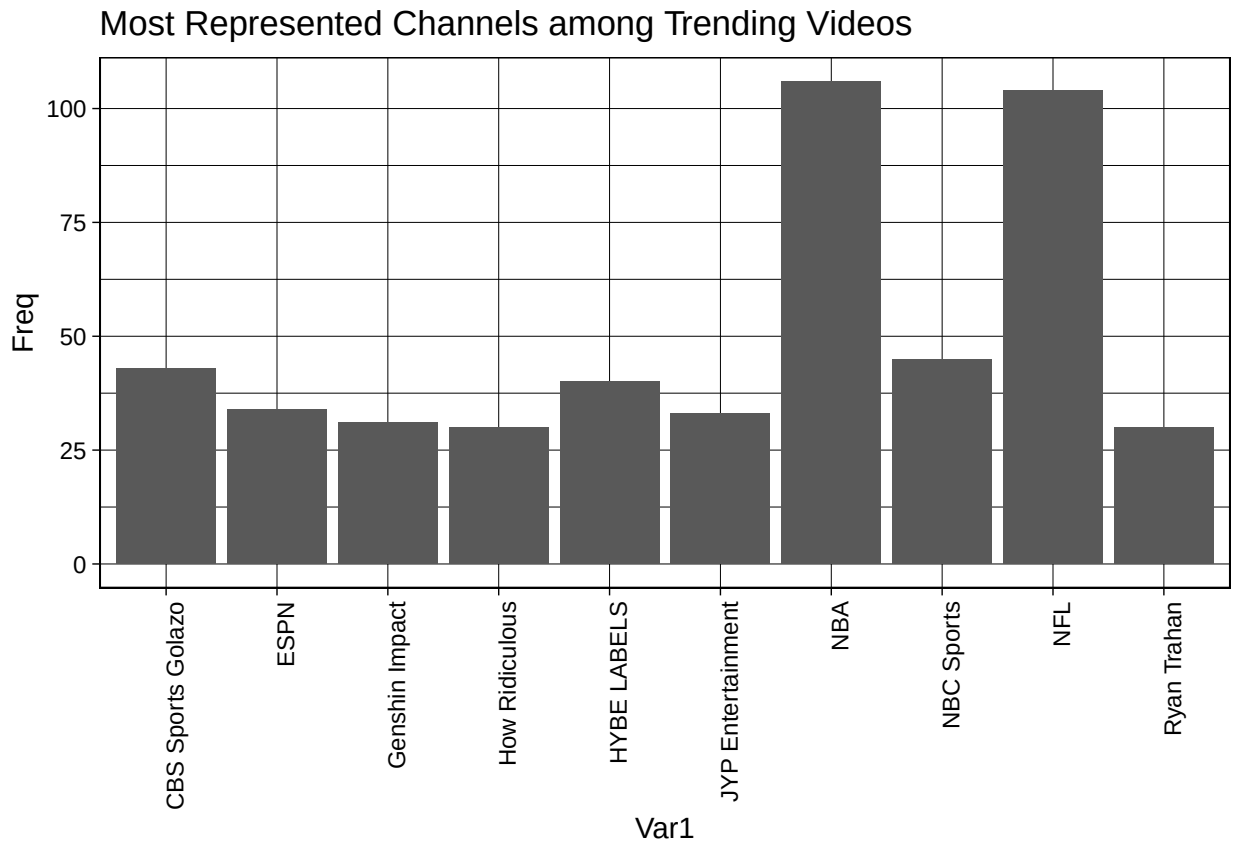


Using Themes

We can spice up our visualizations using another nifty feature of `ggplot`: themes! I've copied and pasted the code from our earlier exercise and added a theme using `theme_linedraw()`. I invite you to go ahead and swap out the theme and replace it with one of the other themes. Examples:

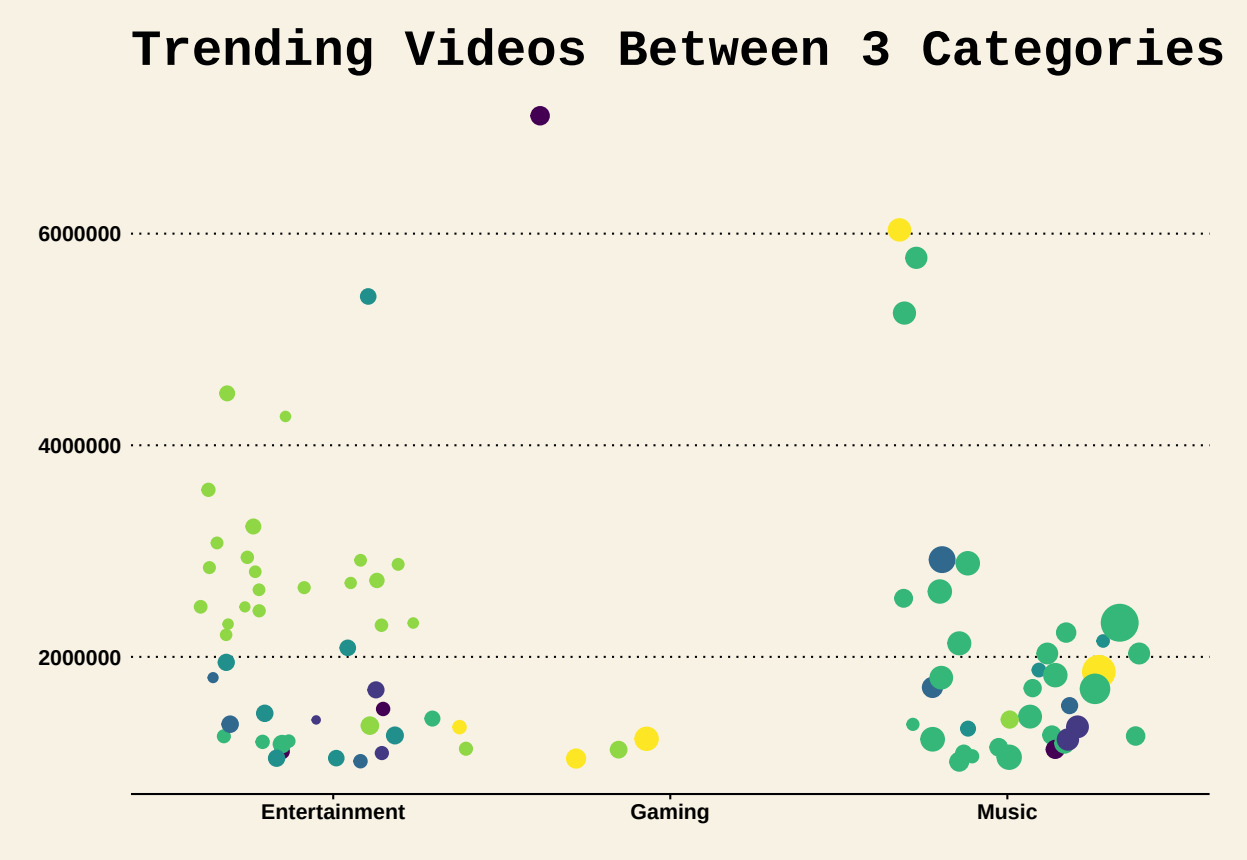
- theme_calc()
- theme_excel() - theme_gdocs()
- theme_classic()

```
ggplot(temp1, aes(x=Var1, y=Freq))+
  geom_col()+
  theme_linedraw()+
  labs(title="Most Represented Channels among Trending Videos")+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Now try one more example - try and replace `theme_wsj` with one of the other themes in the following plot. A simple switch of the overall presentation "theme" using a short function like `theme_` makes it easy to experiment with different aesthetics and is one of yet another benefit of using ggplot!

```
ggplot(data = vids.ags, aes(x=category_id, y=likes, size=comment_count/views))+
  geom_jitter(aes(col=publish_wday))+
  theme_wsj(base_size = 8)+
  labs(title = "Trending Videos Between 3 Categories")+
  theme(legend.position = "none")
```



Introduction to Leaflet [Optional]

Leaflet is among the most popular JS library for interative maps, used by websites such as The New York Times, The Washington Post, GitHub and Flickr[3]. The R package `leaflet` allows us to create leaflet maps directly in R code. The steps are as follow:

1. Create a map widget by calling leaflet().
2. Add layers (i.e., features) to the map by using layer functions (`addTiles`, `addMarkers`, `addPolygons`) to modify the map widget.

Sounds similar enough to the ggplot system? Let's see a simple example.

I'm going to create two objects to be used for our Leaflet map later. First, an icon! Here I'm using Algoritma's main icon and saving it to an object called `ico`. Next, we'll create `loca`, a data frame that has two variables (`lat` and `lng`) with some randomly generated numbers. The code is straightforward:

```
set.seed(418)
library(leaflet)

ico <- makeIcon(
    iconUrl = "https://algorit.ma/wp-content/uploads/2023/04/Algoritma-Logo.png",
    iconWidth=177/2, iconHeight=41/2
```

[3]Official Documentation, Leaflet

```
)


loca <- data.frame(lat=runif(5, min = -6.24, max=-6.23),
                   lng=runif(5, min=106.835, max=106.85))
```

And we'll now create our map:

```
# create a leaflet map widget
map1 <- leaflet()

# add tiles from open street map
map1 <- addTiles(map1)

# add markers
map1 <- addMarkers(map1, data = loca, icon=ico)

map1
```

Supposed we want the end user to be able to click on each of these icons and have a simple pop-up description, we can add that to our map too!

Create the pop-up text:

```
pops <- c(
    "<h3>Algoritma Main HQ</h3><p>Visit us here!</p>",
    "<strong>Algoritma Business Campus</strong>",
    "<h3>In-Construction</h3><p>New Secondary Campus</p>",
    "<strong>Secondary Campus</strong>",
    "<strong>The Basecamp (business-school)</strong>"
)
```

Adding them to our map:

```
map1 <- leaflet()
map1 <- addTiles(map1)
map1 <- addMarkers(map1, data = loca, icon=ico, popup = pops)

map1
```

`leaflet` does a lot more than the simple demonstration above, but since it belongs to the optional part of this coursebook - I'll leave it up to you, the readers, to further explore its possibilities! While I would recommend you to read this references [4] for detailed information about leaflet and to use `ggplot` as the main focus of your graded assignment, I want to leave the choice up to you. Work with your academic mentors to produce a visualization as specified in the learn-by-building module and good luck!

Summary

The coursebook covers many aspects of plotting, including using visualization libraries such as `ggplot2`, `leaflet` and a few other supporting libraries. I hope you've managed to get a good grasp of the plotting philosophy behind `ggplot2`, and have built a few visualizations with it yourself!

[4] Geospatial Data Visualization with R Programming

Happy coding!

Samuel

Annotations

Cleveland (1985), page 264: "Data that can be shown by pie charts always can be shown by a dot chart. This means that judgements of position along a common scale can be made instead of the less accurate angle judgements." This statement is based on the empirical investigations of Cleveland and McGill as well as investigations by perceptual psychologists."