R Notebook

# The Project

Examines between the categories of the videos and their popularity? This includes variables such as: likes, dislikes, comment counts, and total views

# Methodology

Calculate covariance and correlation: This sees the relationships between the variables, and their tendency to change together.

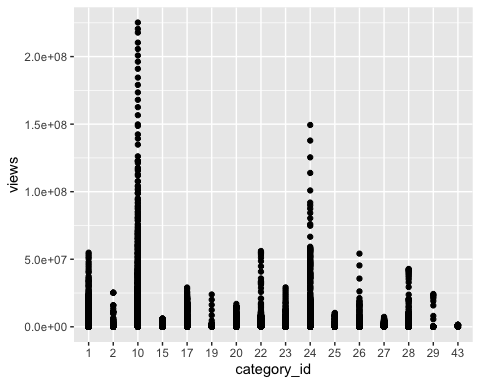
Plot the relationships: This is a visual representation of the relationships of the data.

KNN analysis: This is an analytic approach that is appropriate for predicting a categorical variable. This is fitting for the data because I use the continuous variables to predict the categorical variables.

Logistic Model: This is used to model and predict how many likes a video receives based on what categories the videos are.

library(ggplot2)  
library(class)  
  
df<-read.csv("/Users/muduo/Downloads/ThinkStats2-master/Youtube\_Video\_Analysis/USvideos.csv")  
  
  
df$category\_id <- as.factor(df$category\_id)

ggplot(df, aes(x=category\_id, y=views))+geom\_point()



lm1 <- glm(df$likes ~ df$category\_id)  
summary(lm1)

##   
## Call:  
## glm(formula = df$likes ~ df$category\_id)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -259924 -49306 -28262 -861 5394909   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 70788 4531 15.622 < 2e-16 \*\*\*  
## df$category\_id2 -59731 12080 -4.945 7.66e-07 \*\*\*  
## df$category\_id10 148130 5289 28.007 < 2e-16 \*\*\*  
## df$category\_id15 -49733 8536 -5.826 5.72e-09 \*\*\*  
## df$category\_id17 -25424 6533 -3.892 9.98e-05 \*\*\*  
## df$category\_id19 -58757 11845 -4.960 7.06e-07 \*\*\*  
## df$category\_id20 13714 8915 1.538 0.123956   
## df$category\_id22 -12652 5961 -2.122 0.033806 \*   
## df$category\_id23 -8206 5870 -1.398 0.162186   
## df$category\_id24 -17544 5036 -3.483 0.000495 \*\*\*  
## df$category\_id25 -63490 6316 -10.052 < 2e-16 \*\*\*  
## df$category\_id26 -31502 5670 -5.556 2.78e-08 \*\*\*  
## df$category\_id27 -41043 7044 -5.827 5.68e-09 \*\*\*  
## df$category\_id28 -36414 6371 -5.716 1.10e-08 \*\*\*  
## df$category\_id29 189136 29416 6.430 1.29e-10 \*\*\*  
## df$category\_id43 -51794 29416 -1.761 0.078287 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 48151297662)  
##   
## Null deviance: 2.1452e+15 on 40948 degrees of freedom  
## Residual deviance: 1.9710e+15 on 40933 degrees of freedom  
## AIC: 1123474  
##   
## Number of Fisher Scoring iterations: 2

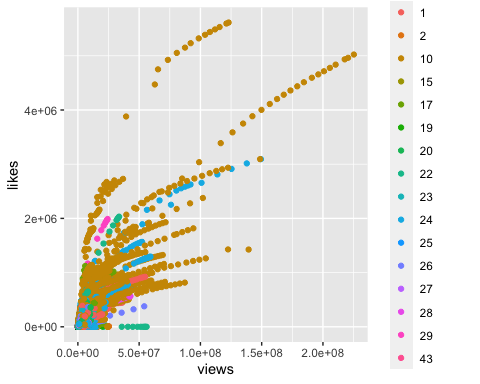
predictions <- predict(lm1, type = "response")  
  
head(predictions)

## 1 2 3 4 5 6   
## 58135.83 53243.33 62582.22 53243.33 53243.33 34374.28

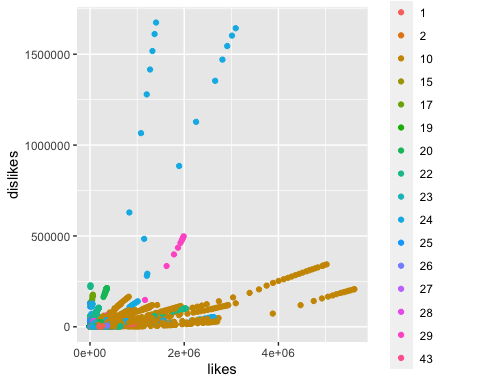
head(df$likes)

## [1] 57527 97185 146033 10172 132235 9763

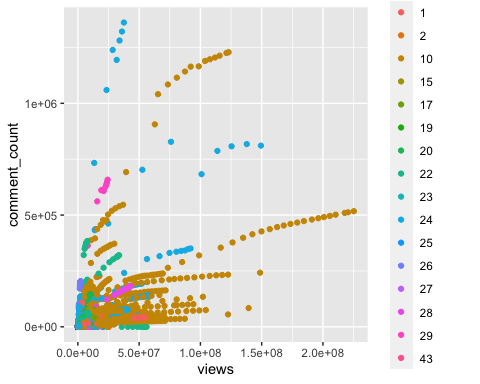
ggplot(data = df, aes(x=views, y=likes, color=category\_id)) + geom\_point()



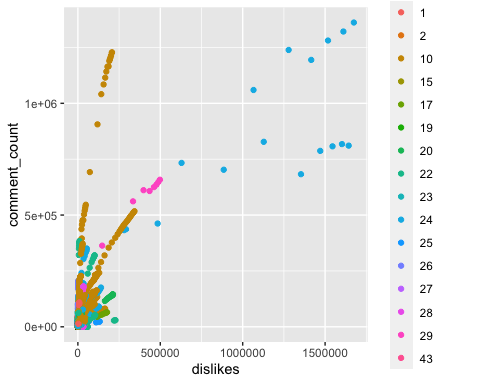
ggplot(data = df, aes(x=likes, y=dislikes, color=category\_id)) + geom\_point()



ggplot(data = df, aes(x=views, y=comment\_count, color=category\_id)) + geom\_point()



ggplot(data = df, aes(x=dislikes, y=comment\_count, color=category\_id)) + geom\_point()



df\_knn <- data.frame(df$likes, df$dislikes, df$views, df$comment\_count, df$category\_id)  
head(df\_knn)

## df.likes df.dislikes df.views df.comment\_count df.category\_id  
## 1 57527 2966 748374 15954 22  
## 2 97185 6146 2418783 12703 24  
## 3 146033 5339 3191434 8181 23  
## 4 10172 666 343168 2146 24  
## 5 132235 1989 2095731 17518 24  
## 6 9763 511 119180 1434 28

classes <- knn(train = df\_knn[1:20001,], test = df\_knn[20002:40002,], cl=df\_knn$df.category\_id[1:20001], k = 29)  
  
accuracy\_table <- data.frame(classes, df\_knn$df.category\_id[1:20001])  
  
head(accuracy\_table)

## classes df\_knn.df.category\_id.1.20001.  
## 1 23 22  
## 2 10 24  
## 3 24 23  
## 4 10 24  
## 5 1 24  
## 6 25 28

sum(accuracy\_table$classes == accuracy\_table$df\_knn.df.category\_id.1.20001.)

## [1] 3408

3421/20000

## [1] 0.17105

# Conclusion

## Implications

It is to be expected that content creating more often than not results in “triviality.” But every once in a while, a video does return high reception, and the video will dominate the trend, and absorb a significant proportion of the view.

## Story of the Data

As expected, the number of views and likes grow with each other. However, there is a difference between the different categories in terms of the number of views, likes, dislikes, and comments the videos receive. With the given data, the model is able to predict the categories correctly 17 percent of the times. We may in turn expect different kinds receptions for the videos based on their genre.

## Limitations of the analysis & Further Development

This analysis could be improved by doing a k-centers cluster analysis. It could also be improved by separating the data into different categories, and see how the ratios between likes, views, dislikes and comments differ.

Another one would be to mine the textual data from the video descriptions, and see if the videos can be separated into their contents or moods or nature of the video. ```