Abhinav Kashyap - 48202300

Mobile Analytics Assignment

#Read data  
geo\_fence <- read.csv("~/Desktop/UCI/Coursework/BANA 277 LEC A- CUST & SOCIAL ANLYTICS/Assignments/Geo-Fence Analytics.csv")  
  
#Creating Dummy variables  
geo\_fence$imp\_large <- ifelse(geo\_fence$imp\_size == '728x90', '1', '0')  
geo\_fence$cat\_entertainment <- ifelse(geo\_fence$app\_topcat == 'IAB1' | geo\_fence$app\_topcat == 'IAB1 - 6','1', '0')  
geo\_fence$cat\_social <- ifelse(geo\_fence$app\_topcat == 'IAB14', '1', '0')  
geo\_fence$cat\_tech <- ifelse(geo\_fence$app\_topcat == 'IAB19-6', '1', '0')  
geo\_fence$os\_ios <-ifelse(geo\_fence$device\_os == 'iOS', '1', '0')  
  
library(aspace)

#Harvesine formula  
geo\_fence$distance <- 6371\*acos(cos(as\_radians(geo\_fence$device\_lat)) \* cos(as\_radians(geo\_fence$geofence\_lat)) \* cos(as\_radians(geo\_fence$device\_lon) - as\_radians(geo\_fence$geofence\_lon)) + sin(as\_radians(geo\_fence$device\_lat)) \* sin(as\_radians(geo\_fence$geofence\_lat)))  
geo\_fence$distance\_squared <- geo\_fence$distance \* geo\_fence$distance   
geo\_fence$ln\_app\_review\_vol <- log(geo\_fence$app\_review\_vol)  
  
colnames(geo\_fence)

## [1] "imp\_size" "app\_id" "app\_name"   
## [4] "app\_pub" "app\_topcat" "app\_review\_vol"   
## [7] "app\_review\_val" "device\_lat" "device\_lon"   
## [10] "device\_zip" "device\_os" "geofence\_lat"   
## [13] "geofence\_lon" "gepfence\_radius" "didclick"   
## [16] "imp\_large" "cat\_entertainment" "cat\_social"   
## [19] "cat\_tech" "os\_ios" "distance"   
## [22] "distance\_squared" "ln\_app\_review\_vol"

#Copying the important variables in a new dataframe for further analysis  
new\_df <- geo\_fence[,c(15,21,22,16,17,18,19,20,23,7)]  
  
  
#Grouping distance into buckets and calculating clickthrough rate  
new\_df$group\_distance = new\_df$distance  
new\_df$group\_distance[new\_df$distance>=0 & new\_df$distance<=0.5] = 1  
new\_df$group\_distance[new\_df$distance>0.5 & new\_df$distance<=1] = 2  
new\_df$group\_distance[new\_df$distance>1 & new\_df$distance<=2] = 3  
new\_df$group\_distance[new\_df$distance>2 & new\_df$distance<=4] = 4  
new\_df$group\_distance[new\_df$distance>4 & new\_df$distance<=7] = 5  
new\_df$group\_distance[new\_df$distance>7 & new\_df$distance<=10] = 6  
new\_df$group\_distance[new\_df$distance>10] = 7  
  
library(stats)  
new\_df$clickthrough\_rate <- ave(new\_df$didclick, new\_df$group\_distance)  
  
#Summarize the new dataframe  
str(new\_df)

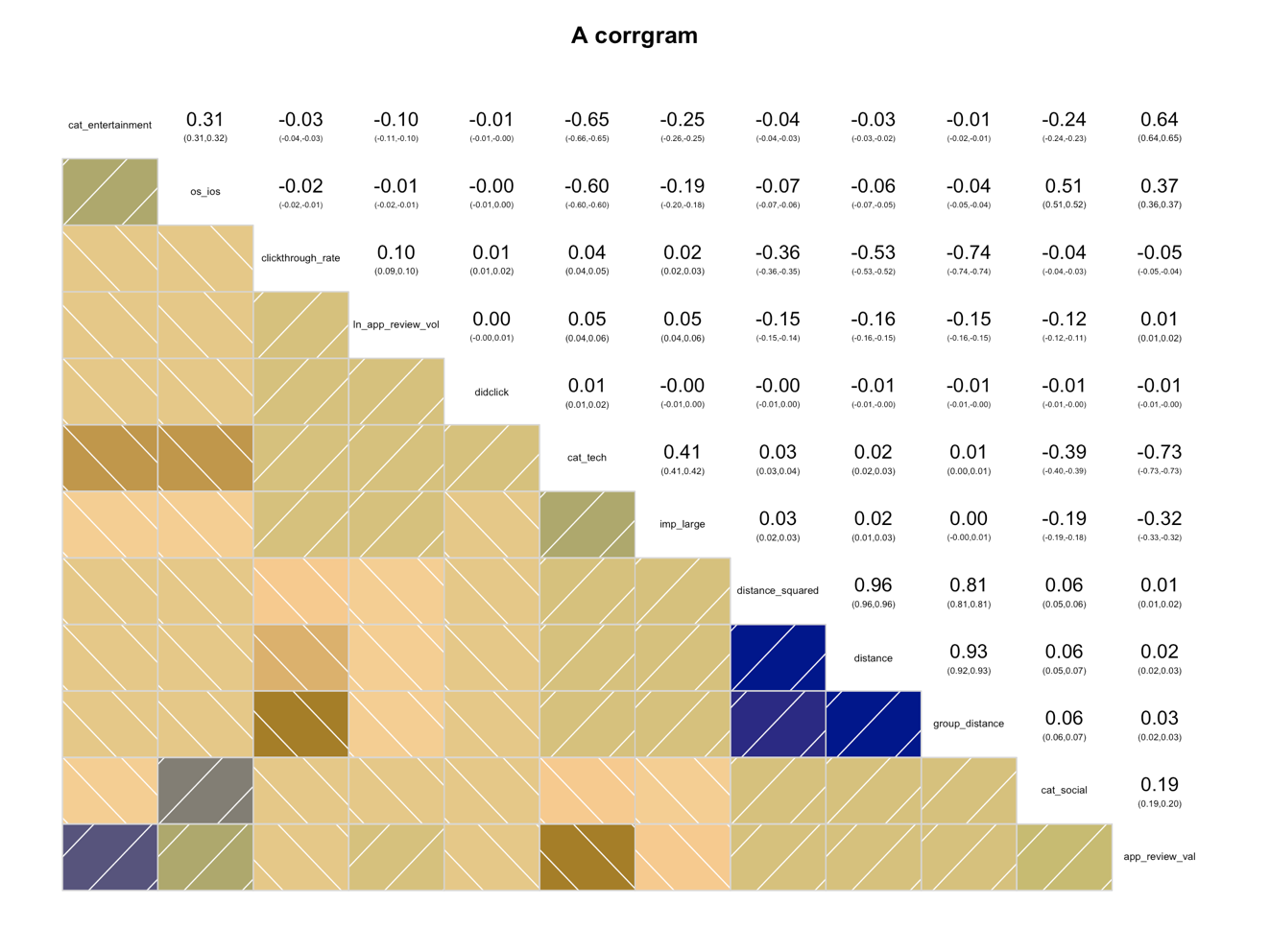
## 'data.frame': 121567 obs. of 12 variables:  
## $ didclick : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ distance : num 2.7 2.43 5.05 2.1 2.1 ...  
## $ distance\_squared : num 7.28 5.88 25.52 4.42 4.39 ...  
## $ imp\_large : chr "0" "0" "0" "0" ...  
## $ cat\_entertainment: chr "0" "0" "0" "0" ...  
## $ cat\_social : chr "0" "0" "0" "0" ...  
## $ cat\_tech : chr "0" "0" "0" "0" ...  
## $ os\_ios : chr "0" "0" "0" "0" ...  
## $ ln\_app\_review\_vol: num 11.1 11.1 11.1 11.1 11.1 ...  
## $ app\_review\_val : num 4.6 4.6 4.6 4.6 4.6 4.6 4.6 4.6 4.6 4.6 ...  
## $ group\_distance : num 4 4 5 4 4 5 4 2 4 4 ...  
## $ clickthrough\_rate: num 0.0058 0.0058 0.00635 0.0058 0.0058 ...

summary(new\_df)

## didclick distance distance\_squared   
## Min. :0.000000 Min. : 0.02076 Min. : 0.00043   
## 1st Qu.:0.000000 1st Qu.: 1.10320 1st Qu.: 1.21705   
## Median :0.000000 Median : 2.02086 Median : 4.08389   
## Mean :0.006811 Mean : 2.98374 Mean : 15.91732   
## 3rd Qu.:0.000000 3rd Qu.: 4.02922 3rd Qu.: 16.23462   
## Max. :1.000000 Max. :11.78666 Max. :138.92542   
## imp\_large cat\_entertainment cat\_social   
## Length:121567 Length:121567 Length:121567   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
## cat\_tech os\_ios ln\_app\_review\_vol app\_review\_val   
## Length:121567 Length:121567 Min. : 7.081 Min. :1.400   
## Class :character Class :character 1st Qu.: 9.791 1st Qu.:3.400   
## Mode :character Mode :character Median :10.087 Median :3.400   
## Mean :10.057 Mean :3.655   
## 3rd Qu.:10.087 3rd Qu.:3.900   
## Max. :12.938 Max. :4.700   
## group\_distance clickthrough\_rate   
## Min. :1.000 Min. :0.005537   
## 1st Qu.:3.000 1st Qu.:0.005801   
## Median :4.000 Median :0.006999   
## Mean :3.622 Mean :0.006811   
## 3rd Qu.:5.000 3rd Qu.:0.007021   
## Max. :7.000 Max. :0.008195

#Converting the catagorical variables to numeric  
new\_df$imp\_large <- as.numeric(new\_df$imp\_large)  
new\_df$cat\_entertainment <- as.numeric(new\_df$cat\_entertainment)  
new\_df$cat\_social <- as.numeric(new\_df$cat\_social)  
new\_df$cat\_tech <- as.numeric(new\_df$cat\_tech)  
new\_df$os\_ios <- as.numeric(new\_df$os\_ios)  
  
#Correlation using library corrgram  
library(corrgram)

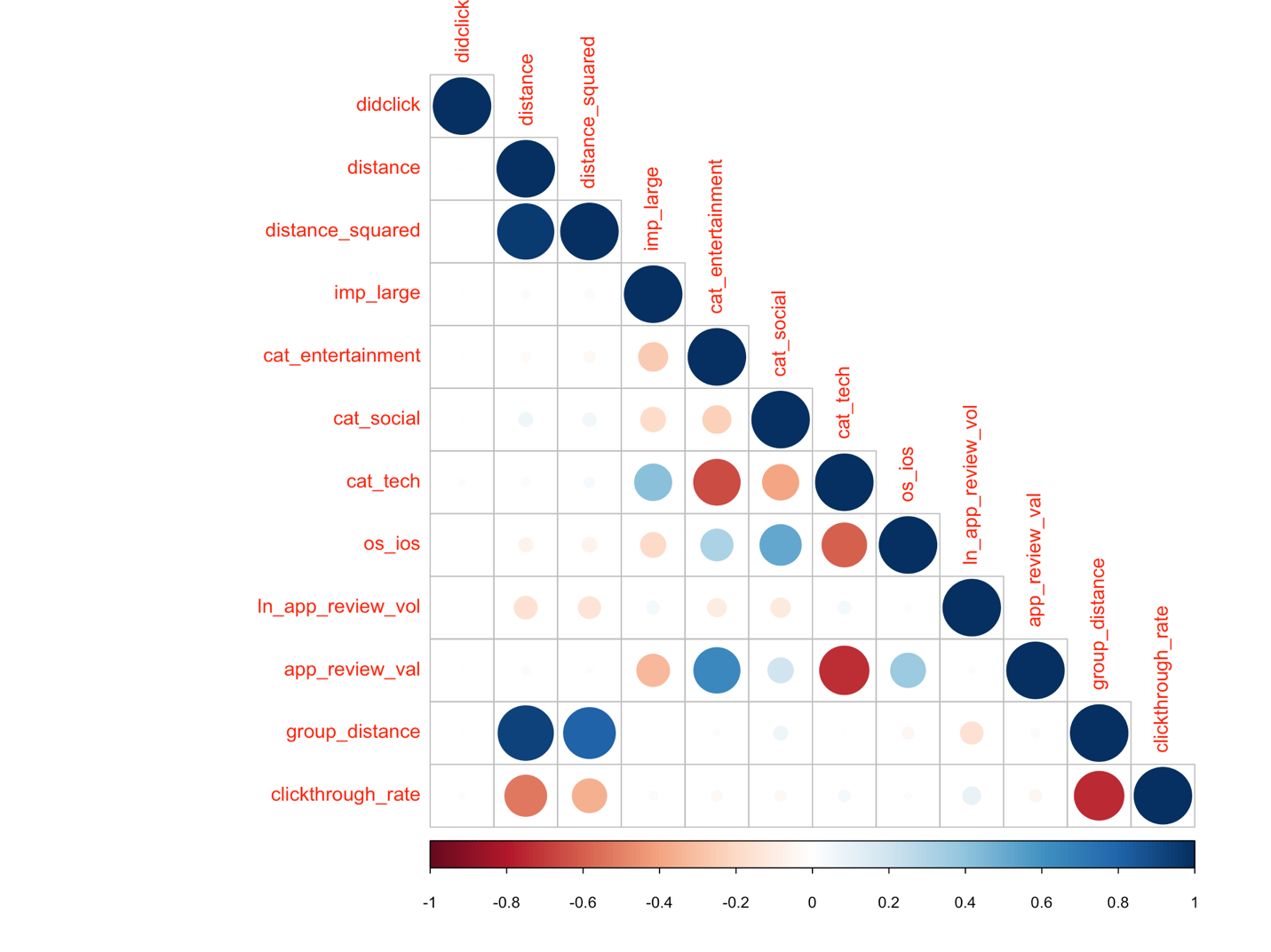
cols <- colorRampPalette(c("darkgoldenrod4", "burlywood1", "darkkhaki", "darkblue"))  
corrgram(new\_df, order = TRUE, col.regions = cols,   
 lower.panel = panel.shade,  
 upper.panel = panel.conf,  
 text.panel = panel.txt,  
 main = 'A corrgram')



#Correlation using library Corrplot  
library(corrplot)

## corrplot 0.84 loaded

correlations <- cor(new\_df)  
corrplot(correlations, method = 'circle', type = 'lower')



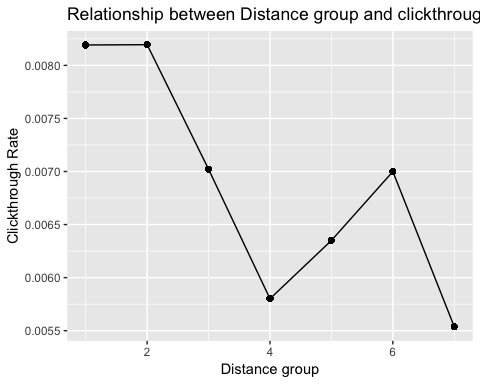
We can observe that there is high positive correlation between

1. app\_review\_val and cat\_entertainment
2. os\_ios and cat\_social

There is high negative correlation between

1. app\_review\_val and cat\_tech
2. group\_distance and clickthrough rate
3. cat\_tech and cat\_entertainment

#plot relationship between distance group and clickthrough rate  
g <- ggplot(new\_df, aes(x=group\_distance, y=clickthrough\_rate)) +   
 geom\_line() + geom\_point() +   
 xlab('Distance group') +   
 ylab('Clickthrough Rate') +  
 ggtitle('Relationship between Distance group and clickthrough rate')  
plot(g)



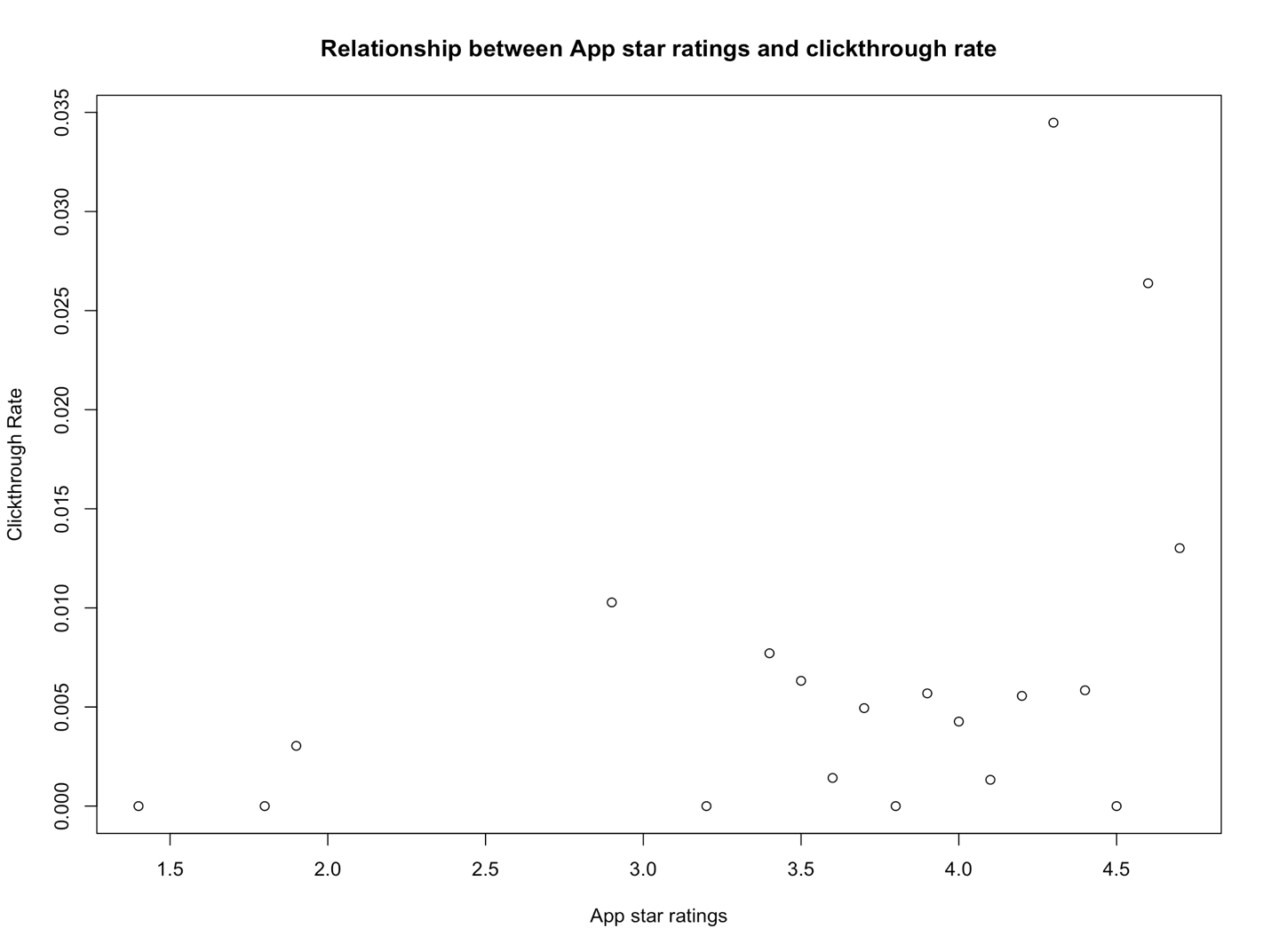
The graph indicates a downward trend for clickthrough rate as the distance increases. This is obvious as people are less likely to click on the advertisement if they are farther away from the main location.

Therefore, distance and clickthrough rate are inversely proportional to some extent. The actual equation coefficient can be found out using Logistic regression (which we’ll see further in this document)

#Some other plots

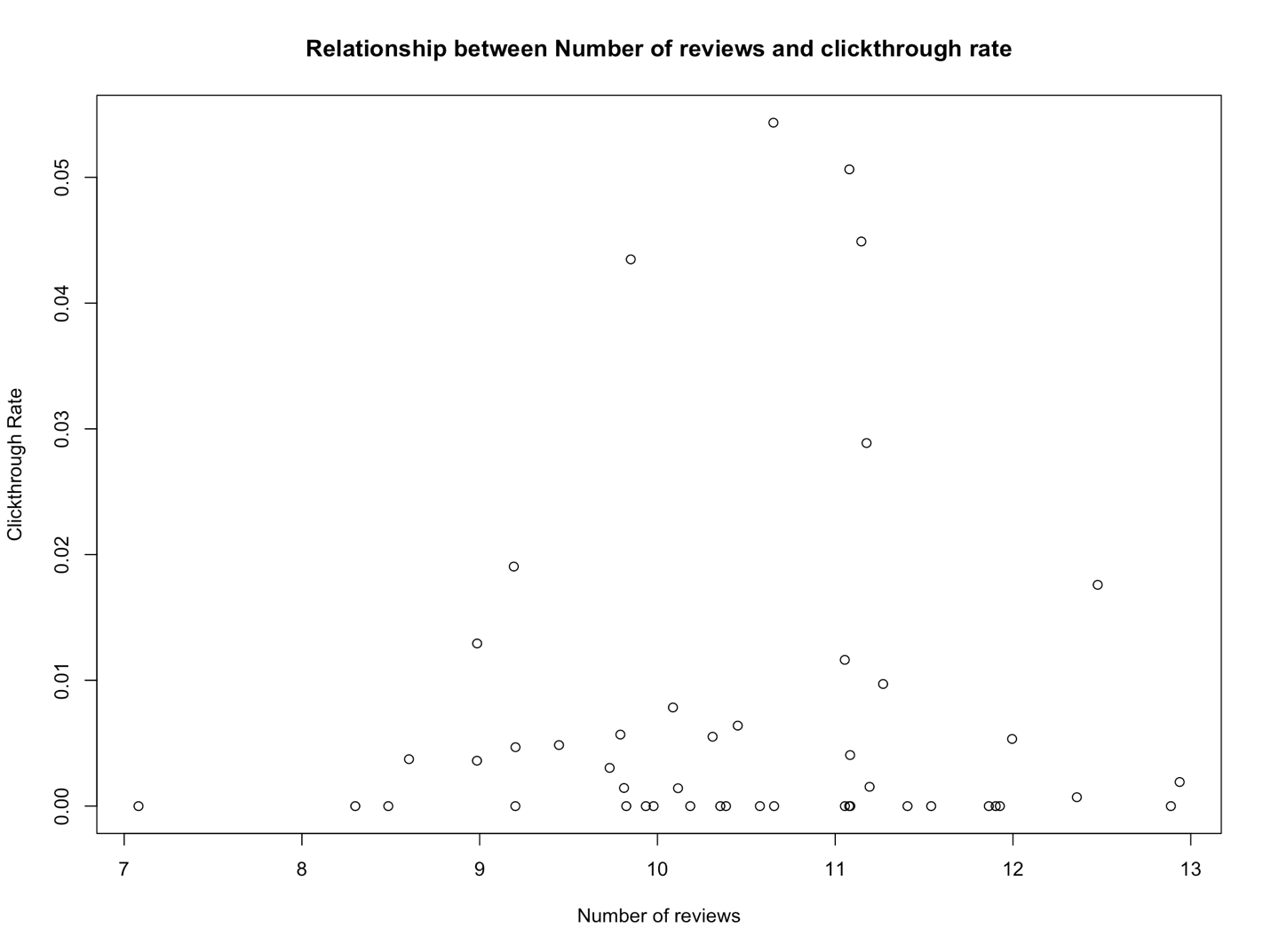
new\_df$impressions <- c(1) #Adding another row to easily sum the number of rows

plot(new\_df %>% group\_by(app\_review\_val) %>% summarise(ctr = sum(didclick)/sum(impressions)), xlab('App star ratings'), ylab('Clickthrough Rate'), title('Relationship between App star ratings and clickthrough rate'))



The app star ratings are usually concentrated from 3 to 5 and most of them have a low click through rate. However, a few of them have a high rating as well as high click through rate.

plot(new\_df %>% group\_by(ln\_app\_review\_vol) %>% summarise(ctr = sum(didclick)/sum(impressions)), xlab('Number of reviews'), ylab('Clickthrough Rate'), title('Relationship between Number of reviews and clickthrough rate'))



The number of reviews is fairly distributed in the dataset and have a random distribution of clickthrough rate.

#Logistic regression  
reg <- glm(didclick ~ distance + distance\_squared + imp\_large + cat\_entertainment +   
 cat\_social + cat\_tech + os\_ios + ln\_app\_review\_vol + app\_review\_val,  
 data = new\_df, family = binomial())  
summary(reg)

##   
## Call:  
## glm(formula = didclick ~ distance + distance\_squared + imp\_large +   
## cat\_entertainment + cat\_social + cat\_tech + os\_ios + ln\_app\_review\_vol +   
## app\_review\_val, family = binomial(), data = new\_df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.1511 -0.1272 -0.1148 -0.1042 3.4022   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -6.620232 0.896730 -7.383 1.55e-13 \*\*\*  
## distance -0.117938 0.045074 -2.617 0.008882 \*\*   
## distance\_squared 0.009167 0.004362 2.102 0.035570 \*   
## imp\_large -0.352155 0.091782 -3.837 0.000125 \*\*\*  
## cat\_entertainment -0.092293 0.178917 -0.516 0.605966   
## cat\_social -0.223206 0.211340 -1.056 0.290900   
## cat\_tech 0.689867 0.176211 3.915 9.04e-05 \*\*\*  
## os\_ios 0.385156 0.126392 3.047 0.002309 \*\*   
## ln\_app\_review\_vol 0.030988 0.062988 0.492 0.622748   
## app\_review\_val 0.322777 0.186744 1.728 0.083908 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 9912.5 on 121566 degrees of freedom  
## Residual deviance: 9857.1 on 121557 degrees of freedom  
## AIC: 9877.1  
##   
## Number of Fisher Scoring iterations: 8

#Checking for colleanearity with VIF  
library(car)

## Loading required package: carData

vif(reg)

## distance distance\_squared imp\_large cat\_entertainment   
## 11.290496 11.259204 1.135594 4.873034   
## cat\_social cat\_tech os\_ios ln\_app\_review\_vol   
## 3.386465 6.162354 2.390056 1.166552   
## app\_review\_val   
## 3.299528

#The VIF value is greater than 10 for distance and distance\_squared indicating high correlation between them. This is obvious as distance squared term has been evaluated from distance. Only keeping distance squared term for our final model  
  
reg1 <- glm(didclick ~ distance\_squared + imp\_large + cat\_entertainment +   
 cat\_social + cat\_tech + os\_ios + ln\_app\_review\_vol + app\_review\_val,  
 data = new\_df, family = binomial())  
summary(reg1)

##   
## Call:  
## glm(formula = didclick ~ distance\_squared + imp\_large + cat\_entertainment +   
## cat\_social + cat\_tech + os\_ios + ln\_app\_review\_vol + app\_review\_val,   
## family = binomial(), data = new\_df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.1413 -0.1328 -0.1133 -0.1048 3.4040   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -6.857181 0.894807 -7.663 1.81e-14 \*\*\*  
## distance\_squared -0.001785 0.001385 -1.289 0.197509   
## imp\_large -0.349254 0.091746 -3.807 0.000141 \*\*\*  
## cat\_entertainment -0.094809 0.179110 -0.529 0.596573   
## cat\_social -0.231336 0.211698 -1.093 0.274496   
## cat\_tech 0.691115 0.176460 3.917 8.98e-05 \*\*\*  
## os\_ios 0.391585 0.126588 3.093 0.001979 \*\*   
## ln\_app\_review\_vol 0.039383 0.062917 0.626 0.531348   
## app\_review\_val 0.316695 0.187306 1.691 0.090876 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 9912.5 on 121566 degrees of freedom  
## Residual deviance: 9863.9 on 121558 degrees of freedom  
## AIC: 9881.9  
##   
## Number of Fisher Scoring iterations: 8

#VIF looks fine now however distance is no longer significant now  
vif(reg1)

## distance\_squared imp\_large cat\_entertainment cat\_social   
## 1.034219 1.134762 4.883834 3.398084   
## cat\_tech os\_ios ln\_app\_review\_vol app\_review\_val   
## 6.180145 2.397616 1.164277 3.312214

#Calculating the exponential of coefficents because log(odds) are difficult to interpret  
exp(coef(reg1))

## (Intercept) distance\_squared imp\_large cat\_entertainment   
## 0.001051875 0.998216724 0.705214108 0.909546241   
## cat\_social cat\_tech os\_ios ln\_app\_review\_vol   
## 0.793472535 1.995939356 1.479323971 1.040168802   
## app\_review\_val   
## 1.372584158

#Model estimation using Anova and chi-squared test  
anova(reg1, test = 'Chisq')

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: didclick  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 121566 9912.5   
## distance\_squared 1 2.5360 121565 9909.9 0.111278   
## imp\_large 1 2.7105 121564 9907.2 0.099687 .   
## cat\_entertainment 1 9.2980 121563 9897.9 0.002294 \*\*  
## cat\_social 1 10.4562 121562 9887.5 0.001222 \*\*  
## cat\_tech 1 10.7375 121561 9876.7 0.001050 \*\*  
## os\_ios 1 8.7846 121560 9868.0 0.003038 \*\*  
## ln\_app\_review\_vol 1 0.9713 121559 9867.0 0.324348   
## app\_review\_val 1 3.0551 121558 9863.9 0.080484 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Difference between null deviance and the residual deviance shows how our model is doing against the null model. The wider this gap the better. 4 variables have high deviance and have a significant p value

*#Testing out another model with just the 3 significant variables observed from last model*  
reg2 <- **glm**(didclick **~** imp\_large **+** cat\_tech **+** os\_ios, data = new\_df, family = **binomial**())  
**summary**(reg2)

##   
## Call:  
## glm(formula = didclick ~ imp\_large + cat\_tech + os\_ios, family = binomial(),   
##     data = new\_df)  
##   
## Deviance Residuals:   
##     Min       1Q   Median       3Q      Max    
## -0.1331  -0.1331  -0.1158  -0.0995   3.2718    
##   
## Coefficients:  
##             Estimate Std. Error z value Pr(>|z|)      
## (Intercept) -5.30555    0.08503 -62.397  < 2e-16 \*\*\*  
## imp\_large   -0.34639    0.09170  -3.777 0.000158 \*\*\*  
## cat\_tech     0.58418    0.10065   5.804 6.48e-09 \*\*\*  
## os\_ios       0.30437    0.11112   2.739 0.006162 \*\*   
## ---  
## Signif. codes:  0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
##     Null deviance: 9912.5  on 121566  degrees of freedom  
## Residual deviance: 9872.2  on 121563  degrees of freedom  
## AIC: 9880.2  
##   
## Number of Fisher Scoring iterations: 8

*This model looks better in terms of significant variables. However, the residual deviance has increased than the previous model.*

*Comparing the 2 models:*

*#Likelihood ratio test to compare both models*  
**library**(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
##     as.Date, as.Date.numeric

**lrtest**(reg1, reg2)

## Likelihood ratio test  
##   
## Model 1: didclick ~ distance\_squared + imp\_large + cat\_entertainment +   
##     cat\_social + cat\_tech + os\_ios + ln\_app\_review\_vol + app\_review\_val  
## Model 2: didclick ~ imp\_large + cat\_tech + os\_ios  
##   #Df  LogLik Df  Chisq Pr(>Chisq)  
## 1   9 -4932.0                       
## 2   4 -4936.1 -5 8.2724     0.1418

*#The 2nd model is doing better due to high Chisq value but is not statistically different than the other model.*

*#Anova means to compare both models*  
**anova**(reg1, reg2)

## Analysis of Deviance Table  
##   
## Model 1: didclick ~ distance\_squared + imp\_large + cat\_entertainment +   
##     cat\_social + cat\_tech + os\_ios + ln\_app\_review\_vol + app\_review\_val  
## Model 2: didclick ~ imp\_large + cat\_tech + os\_ios  
##   Resid. Df Resid. Dev Df Deviance  
## 1    121558     9863.9              
## 2    121563     9872.2 -5  -8.2724

*Final Interpretations:*

We can conclude that the number of clicks depends on the following factors:

1. The size of the impression. The larger the impression size is the less chances the user will click on it. With 1 unit increase in impression size, the clickthrough rate decreases by a factor of 0.7
2. Device type. iOS users click on an ad more often than Android users. Thus, if a user uses an iOS device, the chances of clicking on app increases by a factor of 1.47
3. App category “IAB19-6” is highly significant in predicting higher clickthrough rate. This category was earlier classified as cat\_tech.

We can also find the relationship of other variables in the model with clickthrough rate, although because these variables are not significant in the model, their importance is low.

1. People are less likely to click on the app if they are farther away from desired location.
2. Higher App reviews contribute to higher number of clicks.