Final Draft

Group Five

10/11/2016

Introduction(Marnelia):

Who we are -> digital marketing consultant
What we will do -> help our client, Tinder, keeps growing and find monetization opportunity

Theory : what do you expect to find ?

For our client , we want to use this dataset to solve 2 main problem?

1 What does dating app users look like?

- Is there any characteristic of dating app users?
- Can we use these characteristic to find our potential users?
- Besides demographic data, this dataset also has behavioral data. Can we use these data to improve our app?

2 How can Tinder grow? How can Tinder make money?

- Where to advertise: social media, online shopping website. . . .?
- Other advertisement opportunity? We will use gaming and job data to see if these websites or apps can help us to recruit more users?
- In return, after above strategy, when we have enough user base, we will consider to sell Ads in Tinder app. Can we consider these websites/apps as our potential advertisement buyers? If yes, what kind of data can support our argument?

Method:

1 Describe the dataset

```
cupid_data<-read.csv ("./Dataset/June 10-July 12, 2015 - Gaming, Jobs and Broadband - CSV.csv")
# dimension (row*col)
# how many questions in each section
# how the dataset was collected (date and which research house)
```

Data Cleansing

ASK ALL: EMPLNW Are you now employed full-time, part-time, retired, or are you not employed for pay? {PIAL trend – added 98 'DK' and change REF from '9' to '99'} 1 Employed full-time 2 Employed part-time 3 Retired 4 Not employed for pay 5 (VOL.) Have own business/self-employed 6 (VOL.) Disabled 7 (VOL.) Student 8 (VOL.) Other 98 (VOL.) Don't know 99 (VOL.) Refused

```
table(cupid_data$emplnw)
```

```
##
##   1   2   3   4   5   6   7   8  98  99
## 853 217 601 225  30  51   5   9   3   7
```

```
# The following factors (1,2,3,4) represent 1896/2001 respondents. They are retained while factors (5,6

cupid_data$emplnw_group <-""
for (i in 1:(nrow(cupid_data))){
  if(cupid_data[i,"emplnw"] == "1") cupid_data[i,"emplnw_group"] <- "Employed full-time"
  if(cupid_data[i,"emplnw"] == "2") cupid_data[i,"emplnw_group"] <- "Employed part-time"
  if(cupid_data[i,"emplnw"] == "3") cupid_data[i,"emplnw_group"] <- "Retired"
  if(cupid_data[i,"emplnw"] == "4") cupid_data[i,"emplnw_group"] <- "Not employed for pay"
  if(cupid_data[i,"emplnw"] == "5") cupid_data[i,"emplnw_group"] <- "Others"
  if(cupid_data[i,"emplnw"] == "6") cupid_data[i,"emplnw_group"] <- "Others"
  if(cupid_data[i,"emplnw"] == "7") cupid_data[i,"emplnw_group"] <- "Others"
  if(cupid_data[i,"emplnw"] == "8") cupid_data[i,"emplnw_group"] <- "Others"
  if(cupid_data[i,"emplnw"] == "98") cupid_data[i,"emplnw_group"] <- "Others"
  if(cupid_data[i,"emplnw"] == "99") cupid_data[i,"emplnw_group"] <- "Others"
  }

table(cupid_data$emplnw_group)
```

```
##
##   Employed full-time   Employed part-time Not employed for pay
##                  853                  217                  225
##               Others              Retired
##                  105                  601
```

ASK IF EMPLOYED (EMPLNW=1,2,5): EMPTYPE1 How would you describe the place where you work? [READ] 1 A large corporation 2 A medium-size company 3 A small business 4 A part of the federal, state or local government 5 A school or educational institution, OR 6 A non-profit organization? 7 (VOL.) Other 8 (VOL.) Self-employed/work at home 98 (VOL.) Don't know 99 (VOL.) Refused

```
table(cupid_data$emptype1)
```

```
##
##   1   2   3   4   5   6   7   8  98
## 324 165 266  90 118  83  21  28   5
```

```
# The following factors (1,2,3,4,5) represent 963/1100 respondents for the question. Factor 4 & 5 will

cupid_data$emptype1_group <-""

for (i in 1:(nrow(cupid_data))){
  ifelse(cupid_data[i,"emptype1"] == 1, cupid_data[i, "emptype1_group"] <- "A large corporation",
  ifelse(cupid_data[i,"emptype1"] == 2, cupid_data[i,"emptype1_group"] <- "A medium-size company",
  ifelse(cupid_data[i,"emptype1"] == 3, cupid_data[i,"emptype1_group"] <- "A small business",
  ifelse(cupid_data[i,"emptype1"] == 4, cupid_data[i,"emptype1_group"] <- "Public Sector",
  ifelse(cupid_data[i,"emptype1"] == 5, cupid_data[i,"emptype1_group"] <- "Public Sector",
  ifelse(cupid_data[i,"emptype1"] == 6, cupid_data[i,"emptype1_group"] <- "Others",
  ifelse(cupid_data[i,"emptype1"] == 7, cupid_data[i,"emptype1_group"] <- "Others",
  ifelse(cupid_data[i,"emptype1"] == 8, cupid_data[i,"emptype1_group"] <- "Others",
  ifelse(cupid_data[i,"emptype1"] == 98, cupid_data[i,"emptype1_group"] <- "Others",
  ifelse(cupid_data[i,"emptype1"] == 99, cupid_data[i,"emptype1_group"] <- "Others",
```

```
  cupid_data[i,"emptype1_group"] <- "No data"))))))))))
  }

table(cupid_data$emptype1_group)
```

```
##
##                         A large corporation A medium-size company
##                   901                   324                   165
##      A small business                Others         Public Sector
##                   266                   137                   208
```

ASK ALL: INC Last year – that is in 2014 – what was your total family income from all sources, before taxes? Just stop me when I get to the right category... [READ] {Master INC2} 1 Less than $10,000 2 10 to under $20,000 3 20 to under $30,000 4 30 to under $40,000 5 40 to under $50,000 6 50 to under $75,000 7 75 to under $100,000 8 100 to under $150,000 9 $150,000 or more 98 (VOL.) Don't know 99 (VOL.) Refused

```
table(cupid_data$inc)
```

```
##
##   1   2   3   4   5   6   7   8   9  98  99
## 157 187 180 165 164 275 208 197 163 117 188
```

```
# The team created 4 categories of income level for this analysis.For a sample size of 2001, there are
# Bottom one-third(524 pax) - Between 0 - 30000
# Middle one-third(604) - Between 30000 - 75000
# Upper one-third(405) - Between 75000 - 150000
# Ultra-rich(163) - Above 150000
# Refused/Don't Know(305)

cupid_data$inc_group <-""

for (i in 1:(nrow(cupid_data))){
  if(cupid_data[i,"inc"] == "1") cupid_data[i,"inc_group"] <- "Low-income"
  if(cupid_data[i,"inc"] == "2") cupid_data[i,"inc_group"] <- "Low-income"
  if(cupid_data[i,"inc"] == "3") cupid_data[i,"inc_group"] <- "Low-income"
  if(cupid_data[i,"inc"] == "4") cupid_data[i,"inc_group"] <- "Middle-income"
  if(cupid_data[i,"inc"] == "5") cupid_data[i,"inc_group"] <- "Middle-income"
  if(cupid_data[i,"inc"] == "6") cupid_data[i,"inc_group"] <- "Middle-income"
  if(cupid_data[i,"inc"] == "7") cupid_data[i,"inc_group"] <- "Upper-income"
  if(cupid_data[i,"inc"] == "8") cupid_data[i,"inc_group"] <- "Upper-income"
  if(cupid_data[i,"inc"] == "9") cupid_data[i,"inc_group"] <- "Ultra-rich"
  if(cupid_data[i,"inc"] == "98") cupid_data[i,"inc_group"] <- "Refused/Rejected"
  if(cupid_data[i,"inc"] == "99") cupid_data[i,"inc_group"] <- "Refused/Rejected"
}

table(cupid_data$inc_group)
```

```
##
##       Low-income    Middle-income Refused/Rejected       Ultra-rich
##              524              604              305              163
##     Upper-income
##              405
```

2 Strength

3 Limitation

(Laurence, can you kindly help to fill first part?)

Analysis

1. What does dating app users look like?

1.1 How does our original user look like? demographic analysis (Lawrence & Marnelia) (Yerik)

Income, employment - Lawrence

```
## your code here
```

Age,gender,ideology,..... - Marnelia

```
## your code here
```

Other interesting - Yerik

```
## your code here
```

Insight : summarize a few demographic traits of our original users

1.2 How does our potential users look like? demographic analysis (Marnelia)

Marnelia will use opinion about online dating to group users (eg. anyone who agree with more than three dating opinion will be grouped into potential users.)

```
## your code here
## Find our target users (potential users who are not against online dating but haven't use it before)
```

Insight :summarize a few demographic traits of our potential

2. How to grow and make money?

2.1 The most tradition way of growing is through paid media and collaboration. Can we know more about our users digital profile, such as online shopping, social media, and use this insight to form our marketing strategy? (Mei)

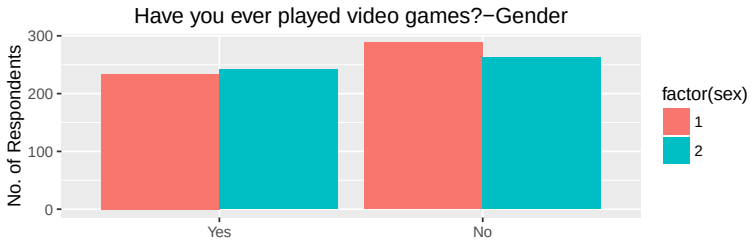
```
## your code here
## the relationship between people's preference or habits of cellphone and Internet in daily life and o
## According to the characteristic of potential users, make our marketing strategy ( where to advertise
```

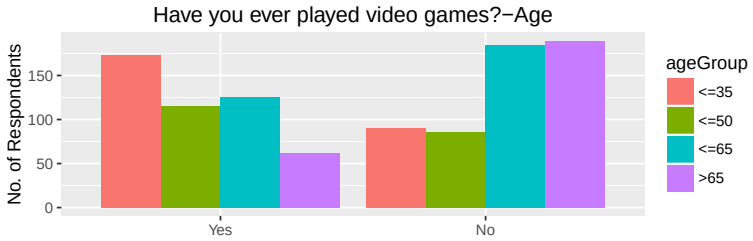
Insight : summarize a few behavior traits of our users and potential users

2.2 Other advertisement opportunity? We will use gaming and job data to see if these websites or apps can help us to recruit more users? (Chewei and Mei-job, Vrat and Yerik-gaming)

```
## your code here -chewei and mei
```

Insight: Can we use gaming/job website as one of growth channel? ######Sample distribution of gaming response across gender and age group





Whether there is a difference in age and gender in terms of gaming in the population?

Chi-squared test is used to test whether there is an inequal distribution in gaming population across gender and age group.

H0: There is no gender difference in the distribution of playing or not playing game in population

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(game1_test)
## X-squared = 0.006646, df = 1, p-value = 0.935
```

p-value is 0.935, which fails to reject H0.

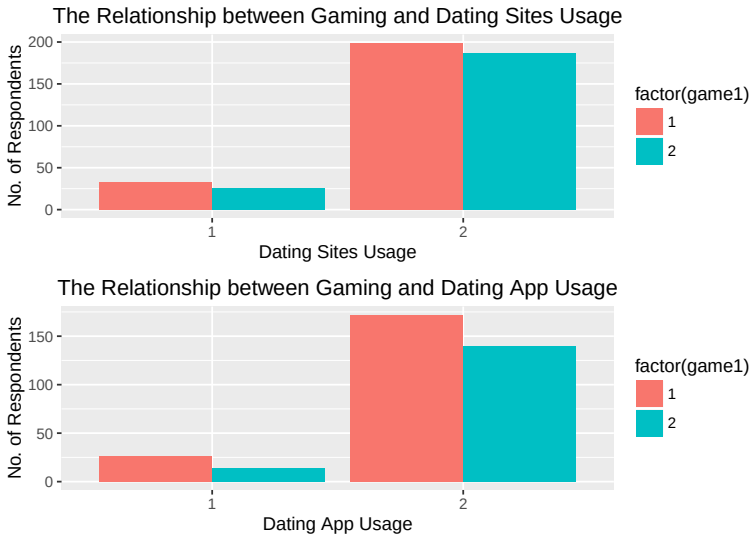
H0: There is no age difference in the distribution of playing or not playing game in population

```
##
##  Pearson's Chi-squared test
##
## data:  table(game1_test)
## X-squared = 42.626, df = 3, p-value = 2.954e-09
```

p-value is 2.954e-09, which rejects H0 at 1% of significance level.

From above two tests, we conclude that there is a significant inequal distribution across age group in gaming population.

Relationship between gaming and online dating



The two charts indicate that among either group who did or did not have online dating experience, more than 50% percent played game before. (how to deduce that this conclusion applies in population?)

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(date1a.game1)
## X-squared = 0.23719, df = 1, p-value = 0.6262
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(date2a.game1)
## X-squared = 1.0315, df = 1, p-value = 0.3098
```

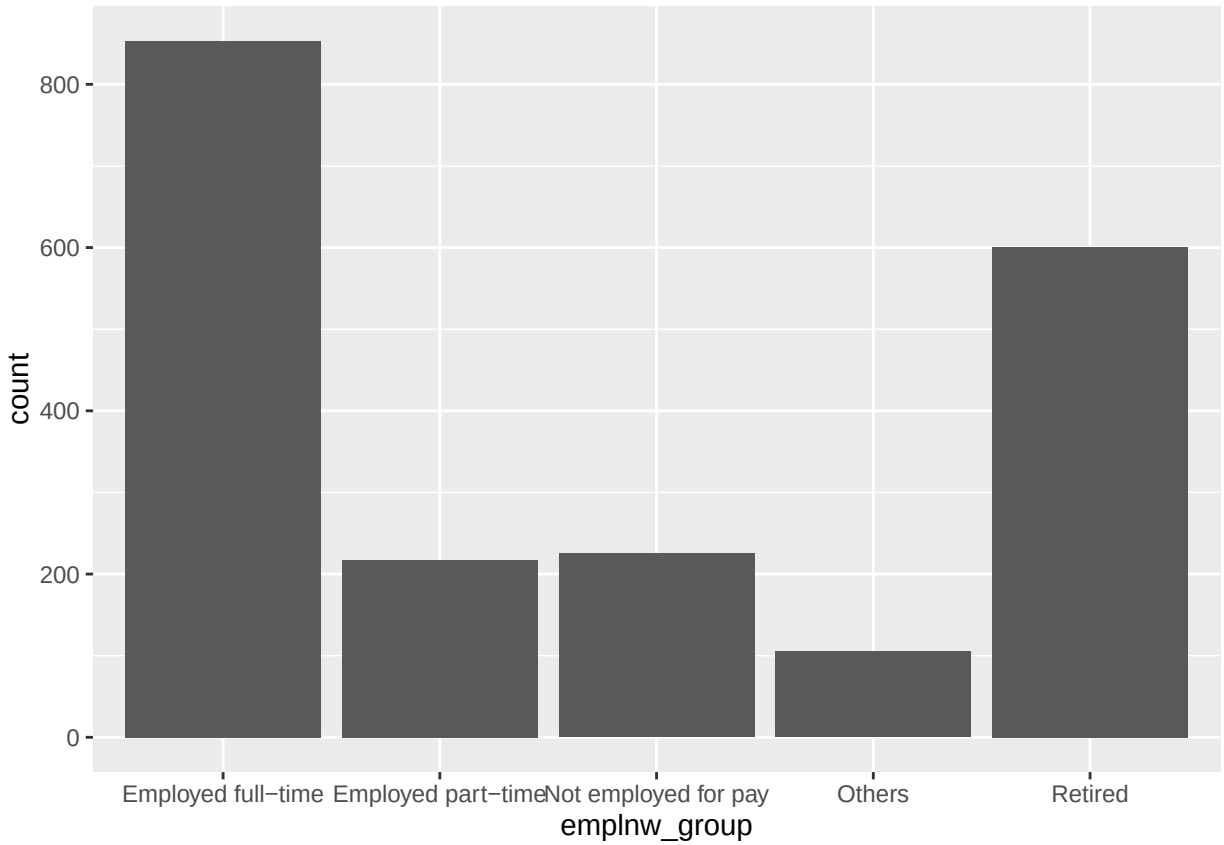
2.3 After the phase of growth, we will seek monetization opportunity. Are the mentioned websites and app our potential advertisement buyers? (Chewei and Mei-job, Vrat and Yerik-gaming)

2.3.1 Difference of online dating between employment status.

Assumption: In this question, we are going to lay the foundatino for our monetization strategy. This part will analyze our users' employment status and if our users are attractive customers of advertisers.

*In this quesiton, we foucs more on people who use dating app.

```
datingpeople1<-datingpeople%>%filter(emplnw==1|emplnw==2|emplnw==3|emplnw==4)
ggplot(data=cupid_data,aes(x=emplnw_group))+geom_bar()
```




```
datejobtable<-table(datingpeople1$emplnw,datingpeople1$date2a)
# exclude emplnw 5~9 (since too little obeservation, and date2a)
datejobtable<-datejobtable[1:4,2:3]
datejobtable
```

```
##
##       1   2
##   1  77 630
##   2  23 136
##   3  13 211
##   4  17 126
```

```
chisq.test(datejobtable)
```

```
##
##  Pearson's Chi-squared test
##
## data:  datejobtable
## X-squared = 8.2935, df = 3, p-value = 0.04032
```

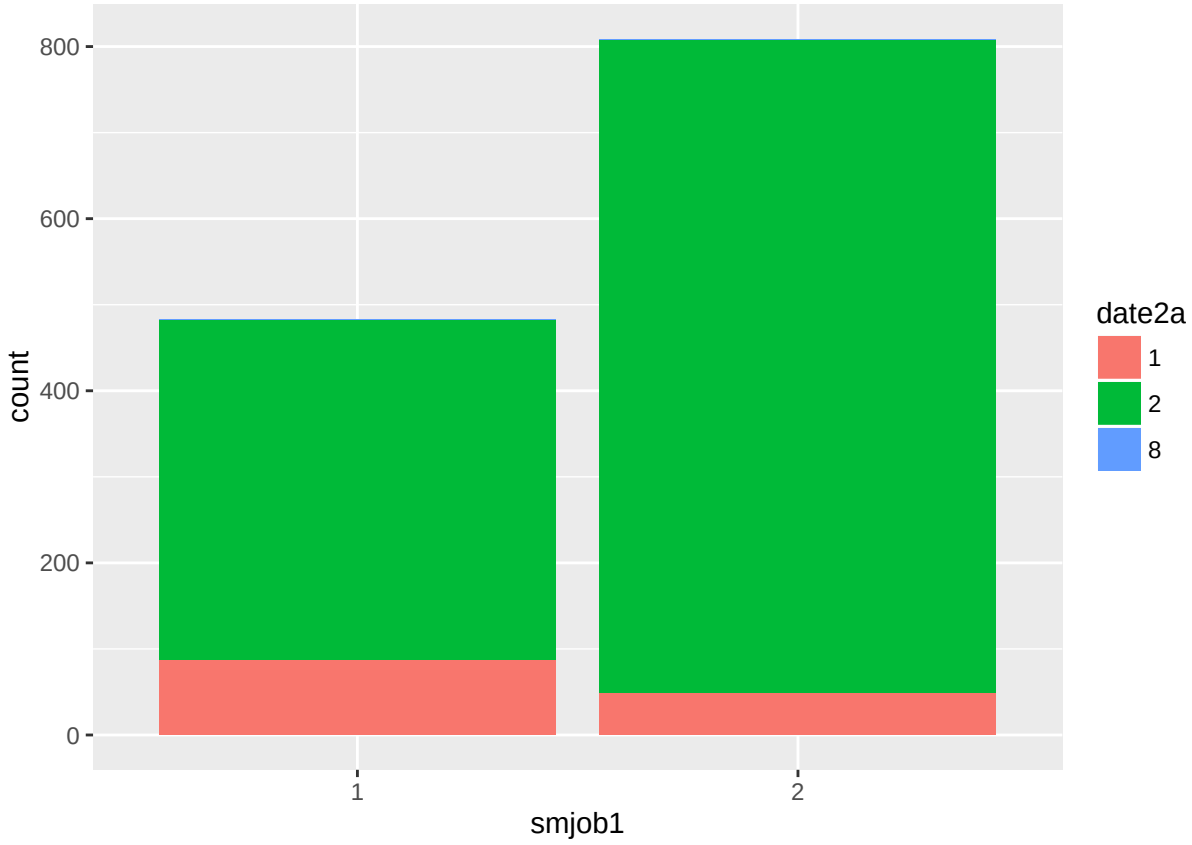
From above graphic, we can find that people who have full-time work are more likely to be a dating app users. We futher conduct a chi-square test to testify if dating app usage will affect their employment status. The p-value is 0.04, below 0.05 siginificant level. Therefore, we prove that our users are potential customers for advertiser.
Action : We will use this figure in our sales kit to convince our advertisers that by investing in our in-app AD, they can acquire huge return.

2.3.2: Is there a model can be used to describe job emplyment and dating app usage?

Assumption: We narrow down our potential customers to online job websites. Here we are going to access if there is any model we can use to predict one's usage of online job seeking.

```
datingpeople[,"age"]<-as.numeric(unlist(datingpeople[,"age"]))
for (i in 1:nrow(datingpeople)){
  if (datingpeople[i,"marital"]=="1"){
    datingpeople[i,"clsmari"]=1
  }
  else if (datingpeople[i,"marital"]=="2"){
    datingpeople[i,"clsmari"]=1
  }
  else{
    datingpeople[i,"clsmari"]=2
  }
}

ggplot(datingpeople, aes(x=smjob1))+geom_bar(aes(fill=date2a))
```




```
dating_between<-glm(smjob1~date2a+age+clsmari,data=datingpeople,family = "binomial")
summary(dating_between)
```

```
##
## Call:
## glm(formula = smjob1 ~ date2a + age + clsmari, family = "binomial",
##     data = datingpeople)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

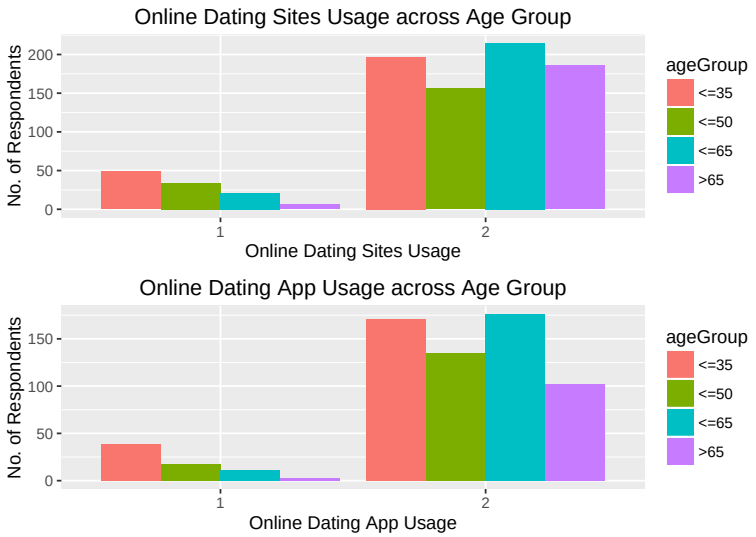
```
## -2.5861  -1.0542   0.5881   0.8739   1.8368
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.721448   0.342723  -5.023 5.09e-07 ***
## date2a2      0.845976   0.210308   4.023 5.76e-05 ***
## date2a8      0.005167   1.444367   0.004    0.997
## age          0.050071   0.004263  11.744  < 2e-16 ***
## clsmari      0.088962   0.139196   0.639    0.523
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1707.9  on 1291  degrees of freedom
## Residual deviance: 1490.2  on 1287  degrees of freedom
## AIC: 1500.2
##
## Number of Fisher Scoring iterations: 4
```

```
pR2(dating_between)
```

```
##          llh      llhNull           G2     McFadden         r2ML
## -745.1249256 -853.9697486  217.6896459    0.1274575    0.1550607
##         r2CU
##    0.2114326
```

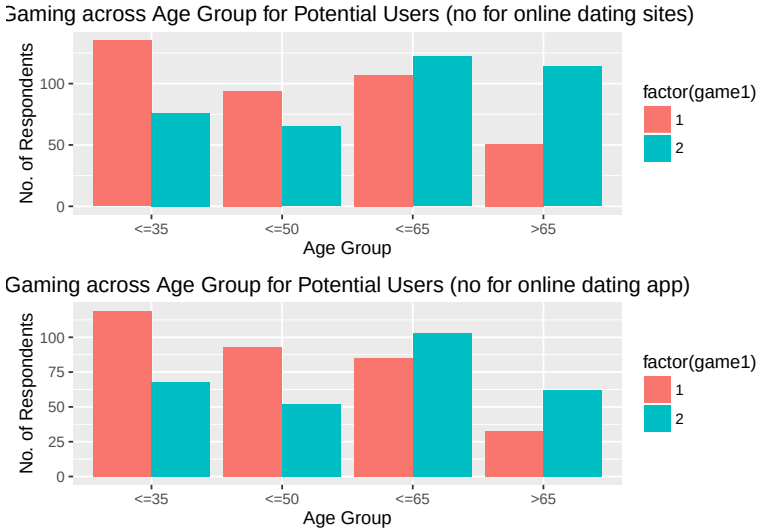
```
#exp(confint(dating_between))
#require(rms)
#mm<-lrm(smjob1~date2a,data=datingpeople)
#summary(mm)
```

If we use date2a (usin dating app) as predictaor of smjob1(using mobile phone to find job), the p-value of intercept, which is the base case of using dating app, is less than 0.05 significance; and p-value of date2a, which is not using dating app, is less than 00.05 significance. #### Action: Merely using dating and age can give us 12% of R^2. We believe that if we can combine more demographic targeting in our tinder database, then we can have a more accurate targeting system for our advertisers.



From the charts, we can see the online dating usage across different age groups. Those under 50 are more

likely to become online dating users. So, will the factor of "gaming" help transform those potential users into users?



For our potential users (people who answered "no" for online dating experience), a large portion played game, espeically for age group under 50. Doing advertisement on video games would be an effective strategy to attract more users for Tinder.

Insight: Can we sell ads to these websites/apps?

Conclusion

1 . Summarize what we find interesting

2 . Next step : What do we suggest Tinder to do?

3 . Further study : What kind of data we want to include in this dataset?