Project 2

Bin Lin

2016-10-9

Project Instruction: Choose any three of the “wide” datasets identified in the Week 6 Discussion items. The goal of this assignment is to give you practice in preparing different datasets for downstream analysis work.

Dataset 1: Religion and Income Distribution Contributor: Yifei Li Source: Introduction to R. (2013). Retrieved from https://ramnathv.github.io (https://ramnathv.github.io)

```
#install.packages("dplyr")
#install.packages("tidyr")
#install.packages("ggplot2")
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

Load the CSV file and transform the data from “wide” to “long”

```
religion_income <- read.csv("C:/Users/blin261/Desktop/DATA607/Religion_Income.csv", header = TRU
E, stringsAsFactors = FALSE, check.names=FALSE)
religion_income
```

```
##   religion <$10k $10-20k $20-30k $30-40k $40-50k $50-75k $75-100k
## 1 Agnostic    27      34      60      81      76     137      122
## 2  Atheist    12      27      37      52      35      70       73
## 3 Buddhist    27      21      30      34      33      58       62
## 4 Catholic   418     617     732     670     638    1116      949
##   $100-150k $>150k
## 1       109     84
## 2        59     74
## 3        39     53
## 4       792    633
```

```
long_data <- religion_income%>%
  gather(income_group, frequency, 2:10)
head(long_data)
```

```
##   religion income_group frequency
## 1 Agnostic        <$10k        27
## 2  Atheist        <$10k        12
## 3 Buddhist        <$10k        27
## 4 Catholic        <$10k       418
## 5 Agnostic      $10-20k        34
## 6  Atheist      $10-20k        27
```

Tidy the data. Get total frequency of income for each individual religion group. I also calculated the percentage of each income group within its religion group.

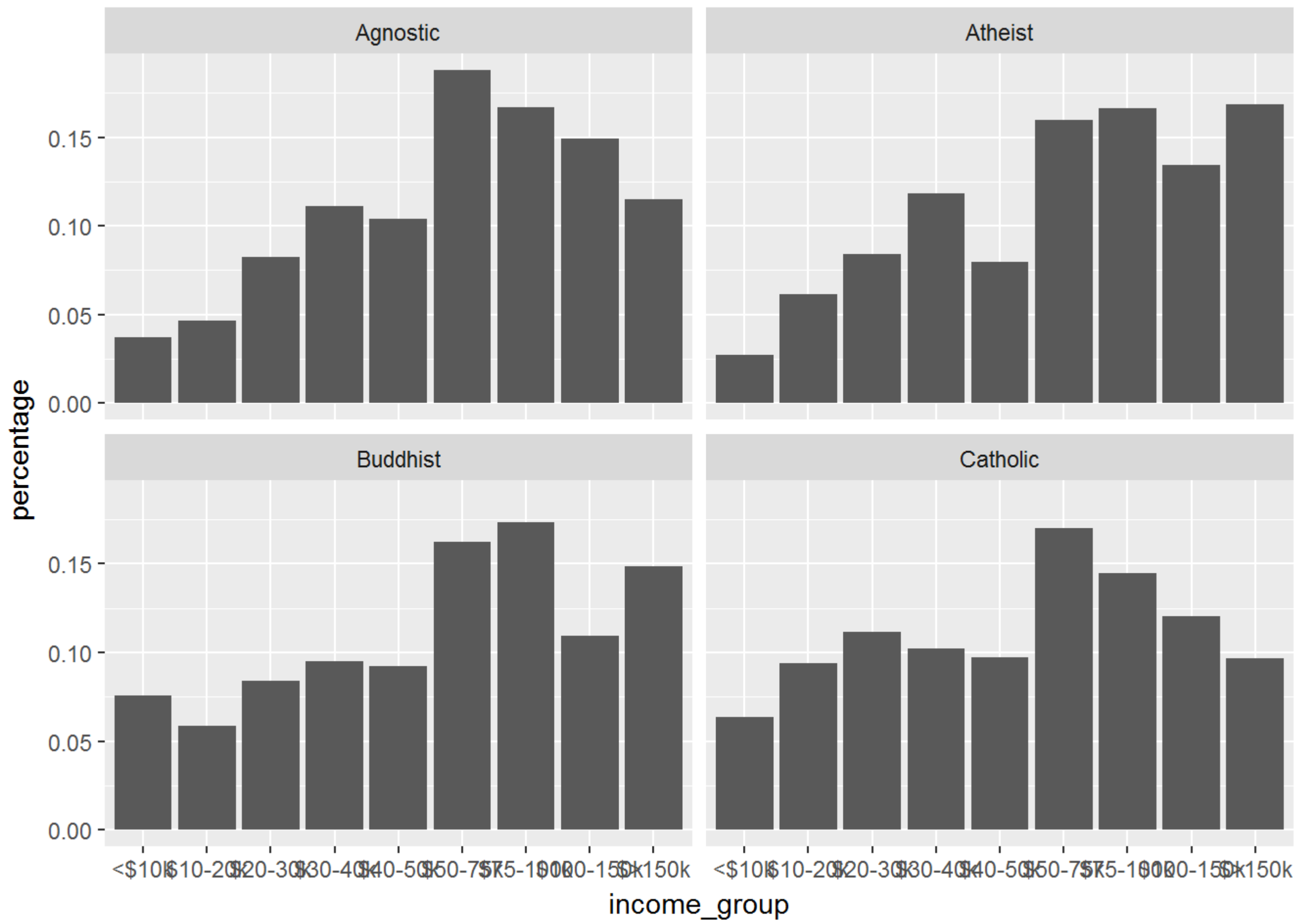
```
r_i <- long_data%>%
  group_by(religion)%>%
  mutate(total = sum(frequency), percentage = frequency/total)%>%
  arrange(religion)
head(r_i)
```

```
## Source: local data frame [6 x 5]
## Groups: religion [1]
##
##   religion income_group frequency total percentage
##      <chr>        <chr>     <int> <int>      <dbl>
## 1 Agnostic        <$10k        27   730 0.03698630
## 2 Agnostic      $10-20k        34   730 0.04657534
## 3 Agnostic      $20-30k        60   730 0.08219178
## 4 Agnostic      $30-40k        81   730 0.11095890
## 5 Agnostic      $40-50k        76   730 0.10410959
## 6 Agnostic      $50-75k       137   730 0.18767123
```

The graph has shown for each religion group, the income distribution normally peaks at $50k-70k, with smallest porportion of people making lower than $10k. As the income keeps going up above $50k-70k, the proportion of people usaully goes down. This makes sense, in real life, we do not see that much people making over 150k.

```
r_i$income_group<-ordered(r_i$income_group,levels=c("<$10k","$10-20k","$20-30k","$30-40k","$40-5
0k","$50-75k","$75-100k", "$100-150k", "$>150k"))

ggplot(data = r_i, aes(x = income_group, y = percentage)) + geom_bar(stat="identity") + facet_wr
ap(~religion)
```



Dataset 2: Gaming Jobs and Broadband Contributor: Bruce Hao Source: http://www.pewinternet.org/datasets/june-10-july-12-2015-gaming-jobs-and-broadband/ (http://www.pewinternet.org/datasets/june-10-july-12-2015-gaming-jobs-and-broadband/)

Load the csv file, and subsetting the variables that will help with our analysis.

```
gaming_job_broadband <- read.csv("C:/Users/blin261/Desktop/DATA607/GamingJobsandBroadband.csv",
header = TRUE, stringsAsFactors = FALSE, check.names=FALSE)

gaming <- gaming_job_broadband %>%
  select(game4, emplnw, stud, age, educ2, inc)
head(gaming)
```

```
##   game4 emplnw stud age educ2 inc
## 1    NA      4    3  47     6  99
## 2     2      3    3  63     4   6
## 3    NA      3    3  86     1   3
## 4    NA      1    3  40     5   6
## 5     2      3    3  65     4   3
## 6    NA      2    3  69     6   8
```

The original data contains observations that are mostly numbers, which stand for certain responses. The following code just making those responses more meaningful by changing the data type from numeric to string which is more human readable. Moreover, it is very helpful to order them in a more sensible sequence which will be easier to perform some analysis later on.

```
gaming$game4[gaming$game4 == 1] <- "gamer"
gaming$game4[gaming$game4 == 2] <- "not_gamer"

gaming$emplnw[gaming$emplnw == 1] <- "full_time"
gaming$emplnw[gaming$emplnw == 2] <- "part_time"
gaming$emplnw[gaming$emplnw == 3] <- "retired"
gaming$emplnw[gaming$emplnw == 4] <- "not_employed"
gaming$emplnw <- ordered(gaming$emplnw, levels = c("full_time", "part_time", "retired", "not_emp
loyed"))


gaming$stud[gaming$stud == 1] <- "full_time_student"
gaming$stud[gaming$stud == 2] <- "part_time_student"
gaming$stud[gaming$stud == 3] <- "no"
gaming$stud <- ordered(gaming$stud, levels = c("full_time_student", "part_time_student", "no"))


gaming$educ2[gaming$educ2 == 1] <- "less_than_HS"
gaming$educ2[gaming$educ2 == 2] <- "HS_incomplete"
gaming$educ2[gaming$educ2 == 3] <- "HS"
gaming$educ2[gaming$educ2 == 4] <- "some_college"
gaming$educ2[gaming$educ2 == 5] <- "associate"
gaming$educ2[gaming$educ2 == 6] <- "bachelor"
gaming$educ2[gaming$educ2 == 7] <- "some_postgraduate"
gaming$educ2[gaming$educ2 == 8] <- "post_graduate"
gaming$educ2 <- ordered(gaming$educ2, levels=c("less_than_HS", "HS_incomplete", "HS", "some_coll
ege", "associate", "bachelor", "some_postgraduate", "post_graduate"))


gaming$inc[gaming$inc == 1] <- "<$10k"
gaming$inc[gaming$inc == 2] <- "$10k-20k"
gaming$inc[gaming$inc == 3] <- "$20-30k"
gaming$inc[gaming$inc == 4] <- "$30-40k"
gaming$inc[gaming$inc == 5] <- "$40k-50k"
gaming$inc[gaming$inc == 6] <- "$50k-75k"
gaming$inc[gaming$inc == 7] <- "$75k-100k"
gaming$inc[gaming$inc == 8] <- "$100k-150k"
gaming$inc[gaming$inc == 9] <- "$>150k"
gaming$inc <- ordered(gaming$inc, levels = c("<$10k", "$10k-20k", "$20k-30k", "$30k-40k", "$40k-
50k", "$50k-75k", "$75k-100k", "$100k-150k", "$>150k"))

head(gaming)
```

```
##       game4       emplnw stud age        educ2        inc
## 1      <NA> not_employed   no  47     bachelor       <NA>
## 2 not_gamer      retired   no  63 some_college   $50k-75k
## 3      <NA>      retired   no  86 less_than_HS       <NA>
## 4      <NA>    full_time   no  40    associate   $50k-75k
## 5 not_gamer      retired   no  65 some_college       <NA>
## 6      <NA>    part_time   no  69     bachelor $100k-150k
```

Still, the data contains observations that does not belong to our interests. We can use functions in dplyr and tidyr to filter out any missing values or response that does not help our analysis.

```
gaming <- gaming%>%
  filter(game4 == "gamer" | game4 == "not_gamer")%>%
  filter(emplnw == "full_time" | emplnw == "part_time" | emplnw ==  "retired" | emplnw == "not_e
mployed")%>%
  filter(stud == "full_time_student" | stud == "part_time_student" | stud == "no")%>%
  filter(educ2 == "less_than_HS" | educ2 == "HS_incomplete" | educ2 == "HS" | educ2 == "some_col
lege" | educ2 == "associate" | educ2 == "bachelor" | educ2 == "some_postgraduate" | educ2 == "po
st_graduate")%>%
  filter(inc == "<$10k" | inc == "$10k-20k" | inc == "$20k-30k" | inc == "$30k-40k" | inc == "$4
0k-50k" | inc == "$50k-75k" | inc == "$75k-100k" | inc == "$100k-150k" | inc == "$>150k")%>%
  arrange(game4, emplnw)
head(gaming)
```

```
##   game4    emplnw              stud age         educ2        inc
## 1 gamer full_time                no  52  some_college $100k-150k
## 2 gamer full_time                no  33 post_graduate   $50k-75k
## 3 gamer full_time part_time_student  61 post_graduate   $50k-75k
## 4 gamer full_time                no  51            HS   $50k-75k
## 5 gamer full_time                no  21      bachelor      <$10k
## 6 gamer full_time                no  26            HS   $10k-20k
```

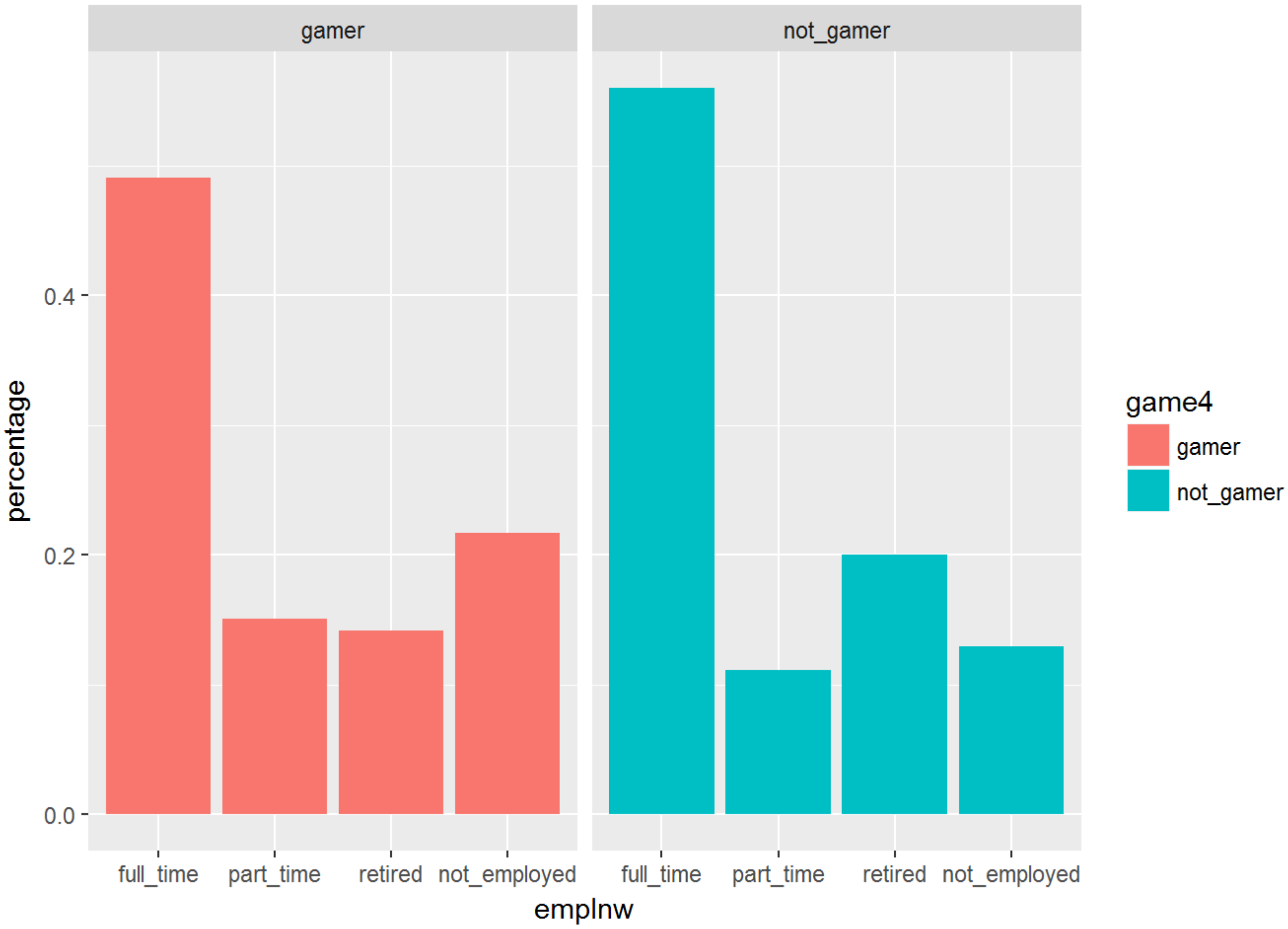
```
View(gaming)
```

The first graph I created compares the relationship between gaming and employment status. We can tell non-gamer has slightly higher percentage of people working full time, and lower percentage of people unemployeed. We also notice there are more people who retired in the not-gamer group. This may be explained by the reason elderly persons may not have quite exposure to internet, computers or smartphones as the young people, therefore, they tend to not playing games.

```
game_emp <- gaming%>%
  group_by(game4, emplnw)%>%
  summarize(count = n())%>%
  mutate(total = sum(count), percentage = count/total)%>%
  arrange(game4,emplnw)
head(game_emp)
```

```
## Source: local data frame [6 x 5]
## Groups: game4 [2]
## 
##       game4       emplnw count total percentage
##       <chr>        <ord> <int> <int>      <dbl>
## 1     gamer    full_time    52   106  0.4905660
## 2     gamer    part_time    16   106  0.1509434
## 3     gamer      retired    15   106  0.1415094
## 4     gamer not_employed    23   106  0.2169811
## 5 not_gamer    full_time   277   495  0.5595960
## 6 not_gamer    part_time    55   495  0.1111111
```

```
ggplot(data = game_emp, aes(x = emplnw, y = percentage, fill = game4)) + geom_bar(stat="identit
y") + facet_wrap(~game4)
```

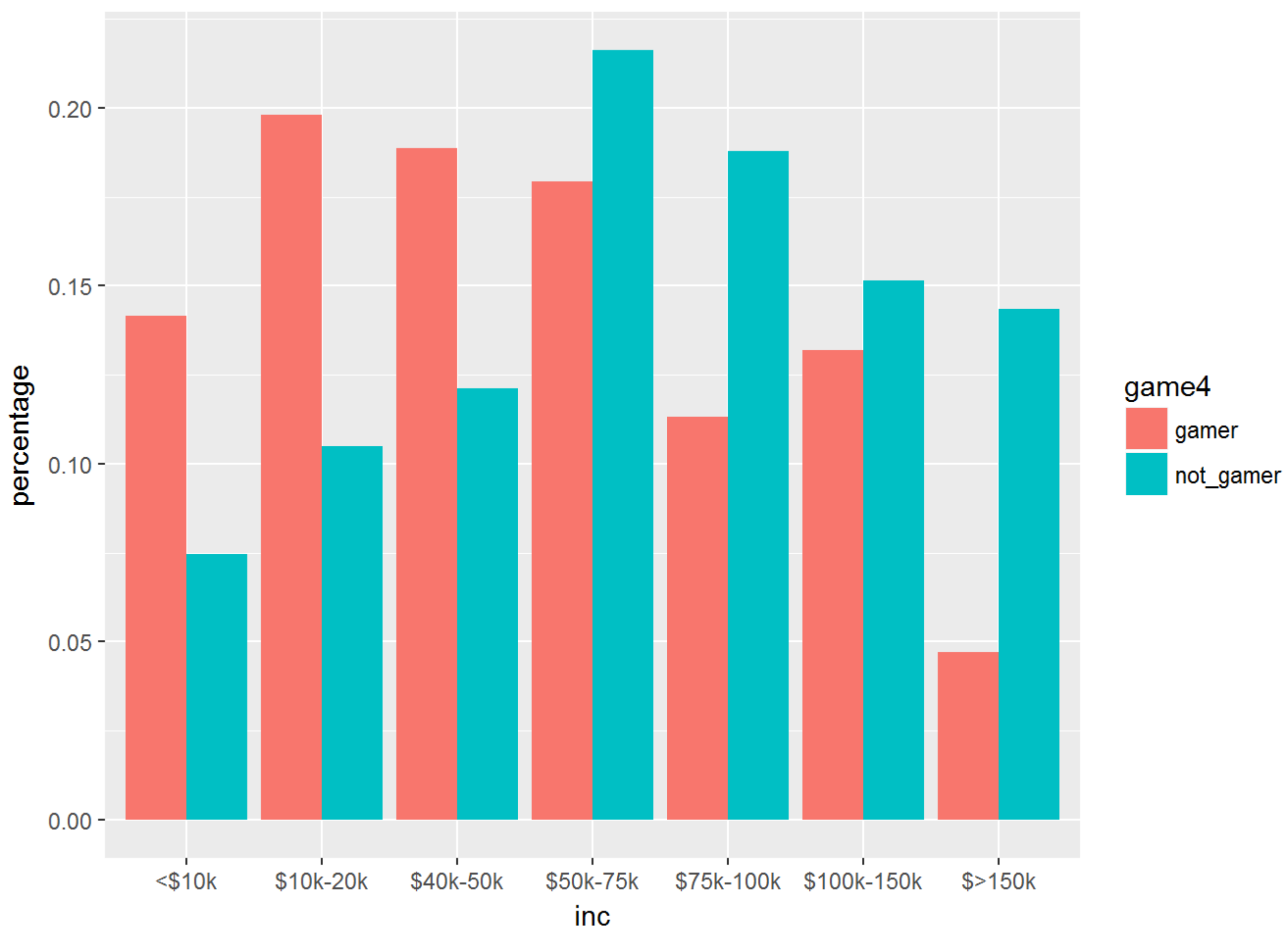


The following graph shows non-gamer makes more money than gamers, as higher proportion of them belong to the higher income group. We can connect this result to the result we got from the first graph. For non-gamers who tend to have full time jobs, of course their income is going to be relatively higher.

```
game_inc <- gaming%>%
  group_by(game4, inc)%>%
  summarize(count = n())%>%
  mutate(total = sum(count), percentage = count/total)%>%
  arrange(game4,inc)
head(game_inc)
```

```
## Source: local data frame [6 x 5]
## Groups: game4 [1]
##
##   game4        inc count total percentage
##   <chr>      <ord> <int> <int>      <dbl>
## 1 gamer      <$10k    15   106  0.1415094
## 2 gamer   $10k-20k    21   106  0.1981132
## 3 gamer   $40k-50k    20   106  0.1886792
## 4 gamer   $50k-75k    19   106  0.1792453
## 5 gamer  $75k-100k    12   106  0.1132075
## 6 gamer $100k-150k    14   106  0.1320755
```

```
ggplot(data = game_inc, aes(x = inc, y = percentage, fill = game4)) + geom_bar(stat="identity",
position = "dodge")
```

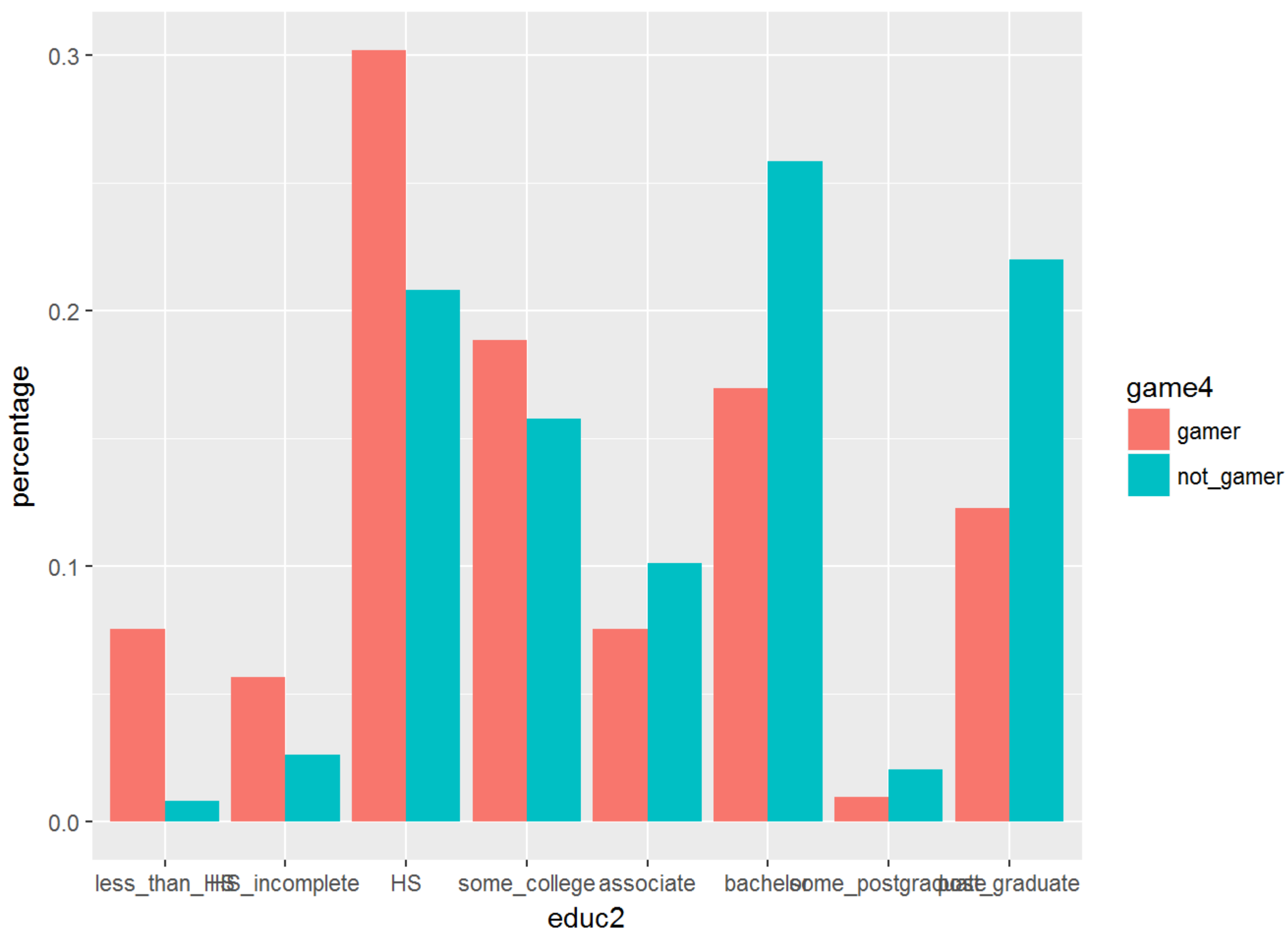


This graph just shows non-gamers have relatively higher education level (obtain a degree higher than high school diploma).

```
game_edu <- gaming%>%
  group_by(game4, educ2)%>%
  summarize(count = n())%>%
  mutate(total = sum(count), percentage = count/total)%>%
  arrange(game4,educ2)
head(game_edu)
```

```
## Source: local data frame [6 x 5]
## Groups: game4 [1]
##
##   game4          educ2 count total percentage
##   <chr>          <ord> <int> <int>      <dbl>
## 1 gamer  less_than_HS     8   106 0.07547170
## 2 gamer HS_incomplete     6   106 0.05660377
## 3 gamer            HS    32   106 0.30188679
## 4 gamer  some_college    20   106 0.18867925
## 5 gamer     associate     8   106 0.07547170
## 6 gamer      bachelor    18   106 0.16981132
```

```
ggplot(data = game_edu, aes(x = educ2, y = percentage, fill = game4)) + geom_bar(stat="identity"
, position = "dodge")
```



Dataset 3: Lending Club Loan Stat 2016Q2 Contributor: Bin Lin Source: https://www.lendingclub.com/info/download-data.action (https://www.lendingclub.com/info/download-data.action)

The first step is to load the data, apparently from the dimention function, we know it is a very large datasets.

```
lending_club <- read.csv("C:/Users/blin261/Desktop/DATA607/LoanStats_2016Q2.csv", header = TRUE,
stringsAsFactors = FALSE)
dim(lending_club)
```

```
## [1] 97856   111
```

Then I tidy, subset, and transform the data. In the meantime, I created a new variable called loantoincome_ratio, which I think is very important variable for us to gain insight about the loan data.

```
loan_stat <- lending_club %>%
  select(term, grade, loan_amnt, annual_inc, int_rate)%>%
  na.omit()
head(loan_stat)
```

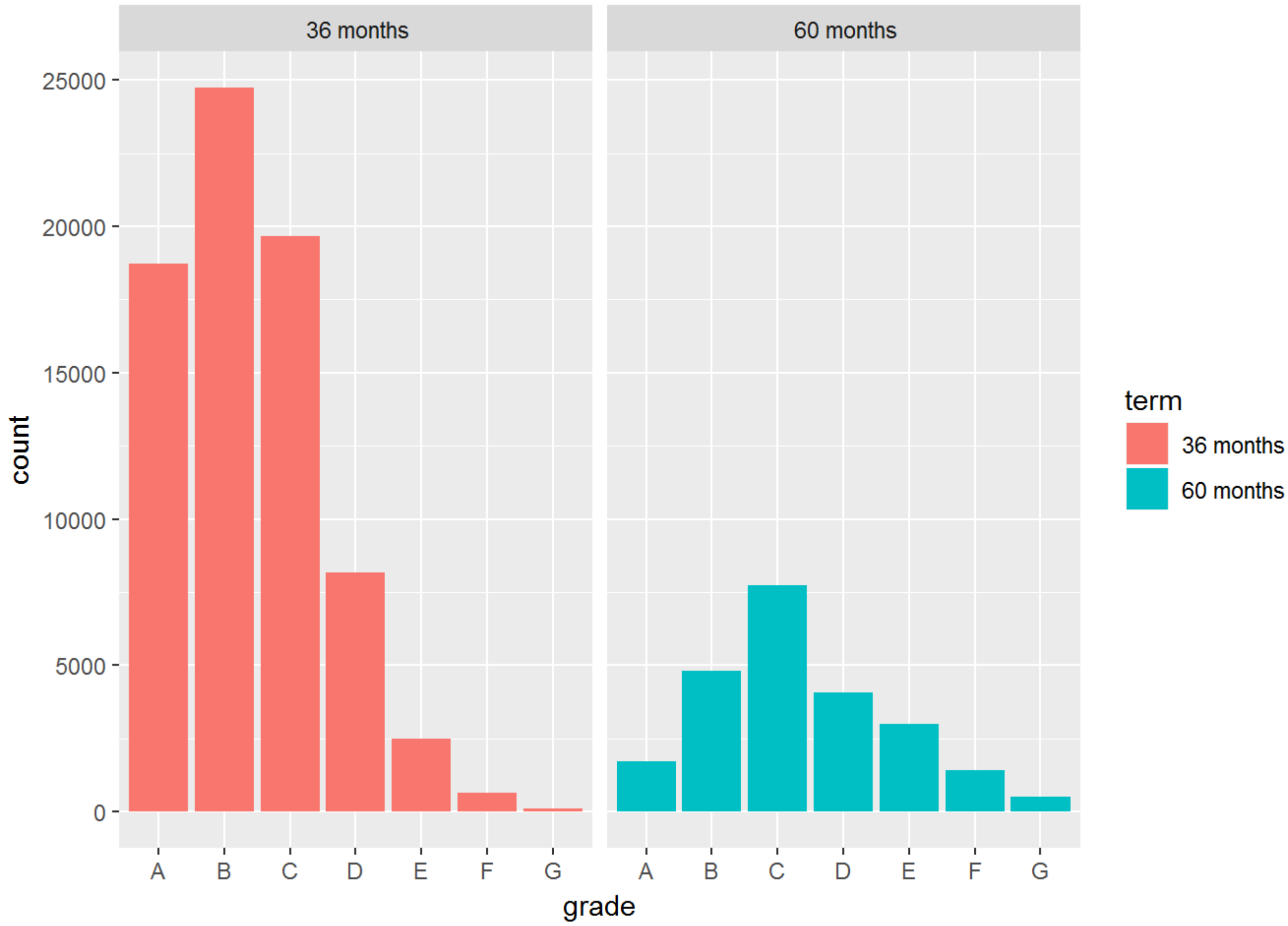
```
##         term grade loan_amnt annual_inc int_rate
## 1  60 months     C     18000      70000   13.49%
## 2  36 months     C      9800      48000   14.49%
## 3  60 months     C     28000      86000   15.59%
## 4  36 months     D     20000      71000   16.99%
## 5  36 months     B      4900     120000   10.99%
## 6  36 months     C     19625      45000   15.59%
```

The first graph shows there are way more 36-month loans approved than the 60-month loans. The distribution are both skewed to the right. The most 36-month loans receive B grade while most 60-month loans receive C grade.

```
loan <- loan_stat %>%
  group_by(term, grade)%>%
  summarize(count = n())
head(loan)
```

```
## Source: local data frame [6 x 3]
## Groups: term [1]
##
##         term grade count
##        <chr> <chr> <int>
## 1  36 months     A 18706
## 2  36 months     B 24729
## 3  36 months     C 19658
## 4  36 months     D  8162
## 5  36 months     E  2510
## 6  36 months     F   651
```

```
ggplot(data = loan, aes(x = grade, y = count, fill = term)) + geom_bar(stat="identity") + facet_
wrap(~term)
```



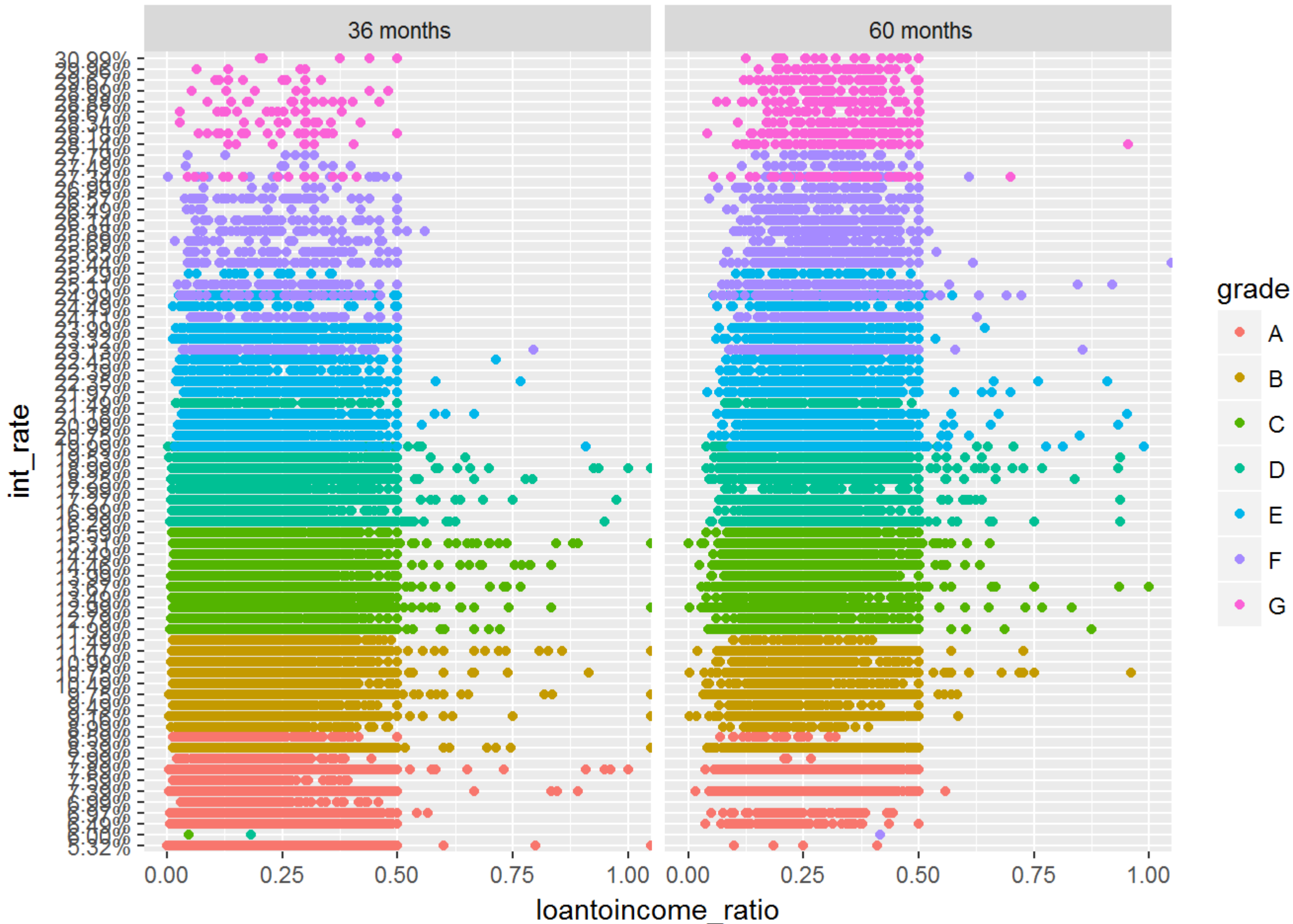
The second graph tell us most of the loans have loan-to-income ratio less than 50%, probabaly because lending club thoughts this type of loan has lower risk. so that the company will be willing to lend the money to these clients. Another thing we found out is on the 60-month loan group, there are more loans with high interest rate (greater than27.34%) and fewer loans with low interest rate (less than 8.59%)

```
loan <- loan_stat %>%
  mutate(loantoincome_ratio = (loan_amnt)/(annual_inc))
head(loan)
```

```
##         term grade loan_amnt annual_inc int_rate loantoincome_ratio
## 1  60 months     C     18000      70000   13.49%         0.25714286
## 2  36 months     C      9800      48000   14.49%         0.20416667
## 3  60 months     C     28000      86000   15.59%         0.32558140
## 4  36 months     D     20000      71000   16.99%         0.28169014
## 5  36 months     B      4900     120000   10.99%         0.04083333
## 6  36 months     C     19625      45000   15.59%         0.43611111
```

```
ggplot(data = loan, aes(x = loantoincome_ratio, y = int_rate, color = grade)) + geom_point(stat=
"identity") + facet_wrap(~term) + xlim(0, 1)
```

```
## Warning: Removed 54 rows containing missing values (geom_point).
```


36 months
60 months
int_rate
30.99%
5.32%
0.00
0.25
0.50
0.75
1.00
loantoincome_ratio
grade
A
B
C
D
E
F
G