Hao-Project2

Bruce Hao

October 3, 2016

```r
library(RCurl)
library(readr)
library(stringr)
library(dplyr)
library(tidyr)
library(ggplot2)
library(ggthemes)
```

Note: This dataset is already in tidy format; however, in order to analyze it, it may be necessary to group and summarize data in various ways. So apologies if this is not a great example of tidying data, but hopefully it is still illustrative of other data grouping and summarization techniques.

Survey questions considered and available answers:

- q3: Overall, how would you rate the economy in your community today? [Excellent, Good, Only fair, OR Poor?]
- date1a: Have YOU, personally, ever used an online dating site such as Match.com, eHarmony, or OK Cupid? [Yes, No]
- game4: Some people use the term "gamer" to describe themselves as a fan of gaming or a frequent game player. Do you think the term "gamer" describes you well, or not? [Yes, gamer, No, not gamer]
- emplnw: Are you now employed full-time, part-time, retired, or are you not employed for pay? [Employed full-time, Employed part-time, Retired, Not employed for pay]
- emtype2: Would you say that the type of work you do primarily involves manual and physical labor, or not? [Yes, involves manual and physical labor, No, does not]
- emptype3: Which of the following best describes the type of work that you do? [Professional, Manager or executive, Government official, Administrative or clerical, OR Customer service?]
- auto2: Thinking about the job or occupation that you work in now, how likely do you think it is that job will exist in its current form in 50 years? Do you think it will definitely exist, probably exist, probably NOT exist, or definitely will NOT exist? [Definitely will exist, Probably will exist, Probably will NOT exist, Definitely will NOT exist…]
- auto3: Overall, how likely do you think it is that in the next 50 years, robots and computers will do much of the work currently done by humans? Do you think this will definitely happen, will probably happen, will probably NOT happen, or will definitely not happen? [Definitely happen, Probably happen, Probably NOT happen, Definitely NOT happen…]
- age: What is your age? [Age]
- educ2: What is the highest level of school you have completed or the highest degree you have received? [Less than high school, High school incomplete, High school graduate, Some college, Two year associate degree, Bachelor degree, Some postgraduate, Postgraduate degree]
- party: In politics TODAY, do you consider yourself a Republican, Democrat, or independent? [Republican, Democrat, Independent]

- inc: Last year – that is in 2014 – what was your total family income from all sources, before taxes [Less than $10,000, 10 to under $20,000, 20 to under $30,000, 30 to under $40,000, 40 to under $50,000, 50 to under $75,000, 75 to under $100,000, 100 to under $150,000, $150,000 or more]

```
url = getURL('https://raw.githubusercontent.com/haobruce/CUNY/master/DATA607/Project2/June%2010-July%20
gamers = read.csv(text=url, stringsAsFactors = F)

# limit columns to only those considered
columns = c('q3', 'date1a', 'game4', 'emplnw', 'emptype2', 'emptype3', 'auto2', 'auto3', 'age', 'educ2'
gamers = gamers[, columns]

# filter rows to include only those that answered the gamer question yes or no
gamers = gamers %>% filter(game4 == 1 | game4 == 2)

# convert answers to string factors
gamers$q3 = as.factor(gamers$q3)
levels(gamers$q3) = list('Excellent'= '1', 'Good' ='2', 'Fair'='3', 'Poor'='4')
gamers$date1a = as.factor(gamers$date1a)
levels(gamers$date1a) = list('Yes'='1', 'No'='2')
gamers$game4 = as.factor(gamers$game4)
levels(gamers$game4) = list('Yes, gamer'='1', 'No, not gamer'='2')
gamers$emplnw = as.factor(gamers$emplnw)
levels(gamers$emplnw) = list('Employed full-time'='1', 'Employed part-time'='2', 'Retired'='3', 'Not emp
gamers$emptype2 = as.factor(gamers$emptype2)
levels(gamers$emptype2) = list('Yes, involves manual and physical labor'='1', 'No, does not'='2')
gamers$emptype3 = as.factor(gamers$emptype3)
levels(gamers$emptype3) = list('Professional'='1', 'Manager or executive'='2', 'Government official'='3
gamers$educ2 = as.factor(gamers$educ2)
levels(gamers$educ2) = list('No high school'='1', 'Some high school'='2', 'High school'='3', 'Some coll
gamers$party = as.factor(gamers$party)
levels(gamers$party) = list('Republican'='1', 'Democrat'='2', 'Independent'='3')
gamers$inc = as.factor(gamers$inc)
levels(gamers$inc) = list('<$10K'='1', '$10-20K'='2', '$20-$30K'='3', '$30K-$40K'='4', '$40K-$50k'='5',
```

Since I will have to copy and paste the plotting and prop.text functions multiple times, I'll write a simple function to avoid copying and pasting and the potential errors that might be introduced in the process.

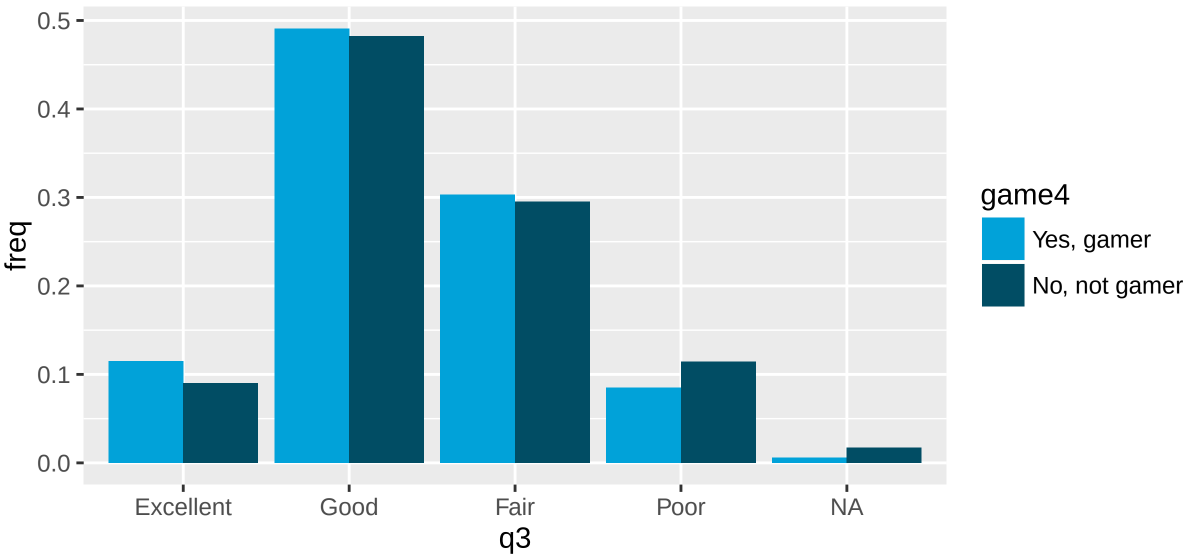
```
plotGamers = function(data, x, group, freq) {
  p = data %>%
    group_by_(group, x) %>%
    summarise(n = n()) %>%
    mutate(freq = n / sum(n)) %>%
    ggplot(aes_string(x=x, y=freq, fill=group)) +
    geom_bar(stat = 'identity', position = 'dodge') +
    scale_x_discrete(labels = function(x) str_wrap(x, width = 10)) +  # wraps x-axis labels
    scale_fill_economist()

  list(p, prop.test(table(gamers[,x], gamers[,group])))
}
```

I wanted to see if self-identified gamers answered certain quesetions different. So starting from the top…

```r
plotGamers(gamers, 'q3', 'game4', 'freq')
```

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



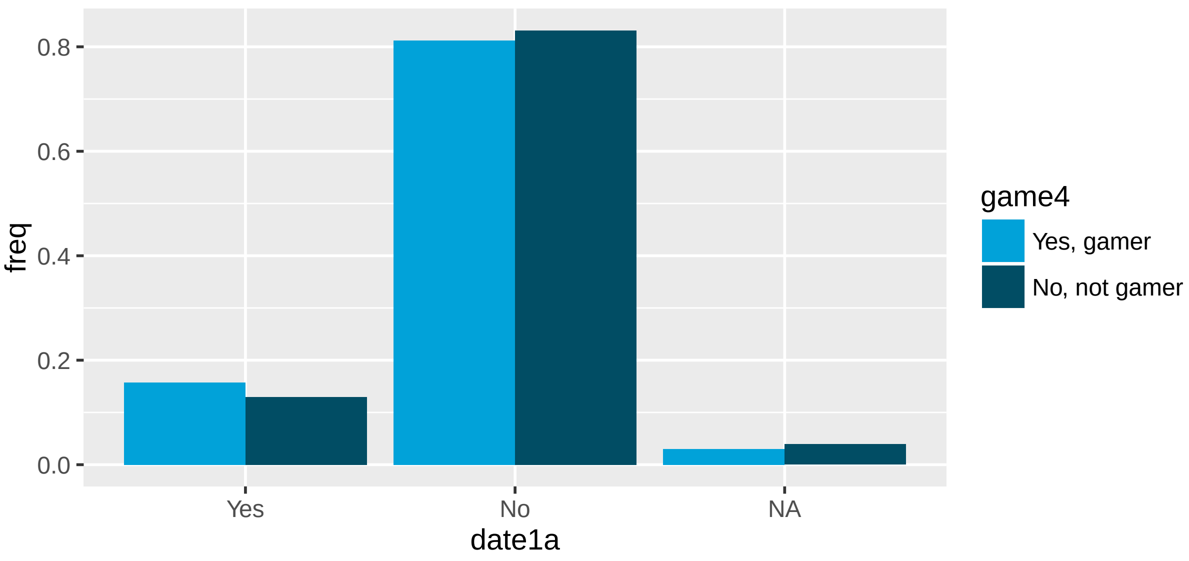
```
##
## [[2]]
##
##  4-sample test for equality of proportions without continuity
##  correction
##
## data:  table(gamers[, x], gamers[, group])
## X-squared = 1.9862, df = 3, p-value = 0.5753
## alternative hypothesis: two.sided
## sample estimates:
##    prop 1    prop 2    prop 3    prop 4
## 0.2209302 0.1845103 0.1858736 0.1414141
```

Based on the chart and 4-sample proportions test, there is no evidence suggesting that gamers view the economy any different than non-gamers.

Next, let's look at dating…

```r
plotGamers(gamers, 'date1a', 'game4', 'freq')
```

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



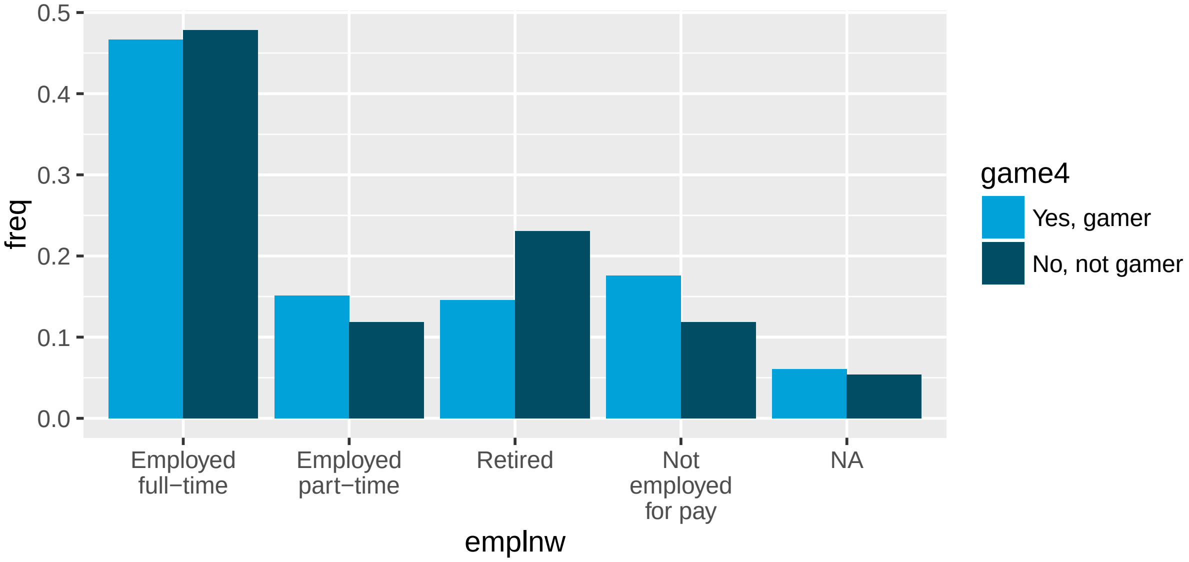
```
##
## [[2]]
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  table(gamers[, x], gamers[, group])
## X-squared = 0.62774, df = 1, p-value = 0.4282
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.04773231  0.11710429
## sample estimates:
##    prop 1    prop 2
## 0.2131148 0.1784288
```

Again, there is no evidence suggesting that gamers using dating sites/apps any more or less than non-gamers.

Next, let's examine employment…

```
plotGamers(gamers, 'emplnw', 'game4', 'freq')
```

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



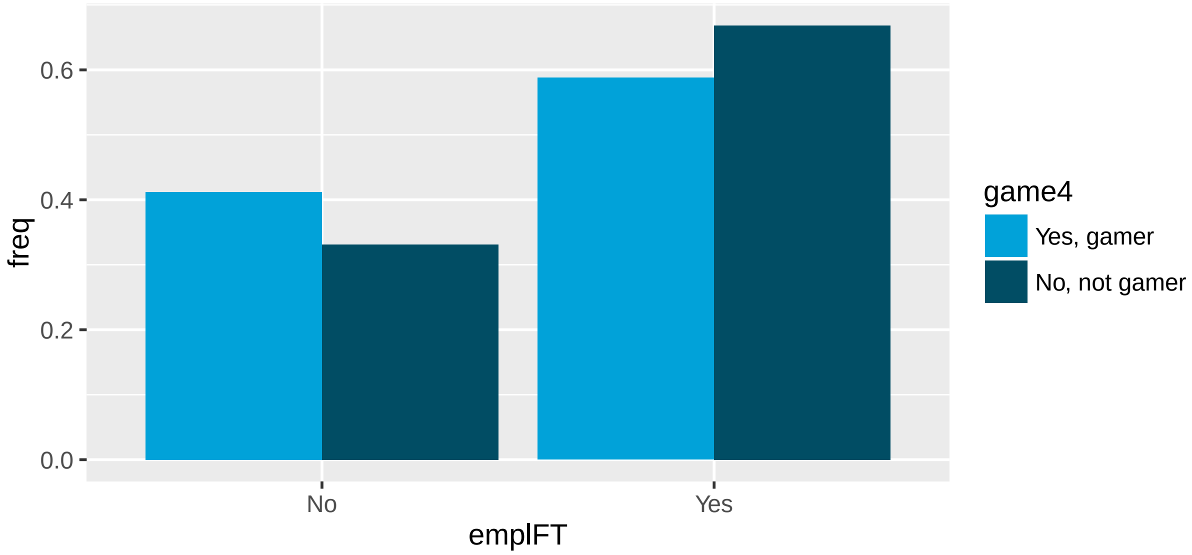
```
##
## [[2]]
##
##  4-sample test for equality of proportions without continuity
##  correction
##
## data:  table(gamers[, x], gamers[, group])
## X-squared = 9.2037, df = 3, p-value = 0.0267
## alternative hypothesis: two.sided
## sample estimates:
##    prop 1    prop 2    prop 3    prop 4
## 0.1782407 0.2212389 0.1230769 0.2478632
```

Optically, it appears that gamers are slightly less likely to be employed full-time and much less likely to be retired (which makes sense intuitively). Conversely, gamers are more likely to be employed part-time or not employed at all. The 4-sample proportion equality test has a p-value less than 0.05, which suggests that the proportion of gamers within each employment category are statistically different from one another.

Let's limit the sample further to only working age adults by removing retiress; furthermore, let's consolidate employment status to just employed full-time or not and see if there is a difference then.

```
NotRetired = gamers %>% filter(emplnw != 'Retired') %>%
  mutate(emplFT = ifelse(emplnw == 'Employed full-time', 'Yes', 'No'))

NotRetired %>%
  group_by(game4, emplFT) %>%
  summarise(n = n()) %>%
  mutate(freq = n / sum(n)) %>%
  ggplot(aes(x=emplFT, y=freq, fill=game4)) +
  geom_bar(stat = 'identity', position = 'dodge') +
  scale_fill_economist()
```



```
prop.test(table(NotRetired$emplFT, NotRetired$game4))
```

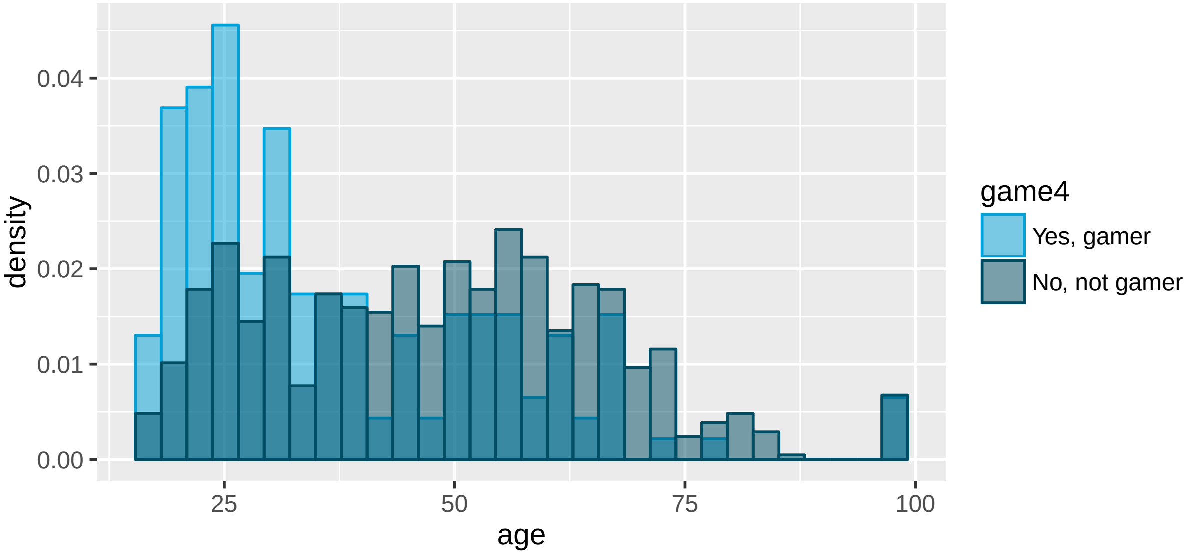
```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  table(NotRetired$emplFT, NotRetired$game4)
## X-squared = 2.6773, df = 1, p-value = 0.1018
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.01238774  0.12547147
## sample estimates:
##    prop 1    prop 2
## 0.2347826 0.1782407
```

In this particular case, optically it still appears that gamers are less likely to be employed full-time, but the 2-sample test has a p-value of 0.10 which suggests that this difference may be due simply to chance.

Next, let’s take a look at age. Since age is captured as a number (not a factor of bins), we’ll use a histogram to have a look…

```
gamers %>%
  ggplot(aes(x=age, color=game4, fill=game4), addDensity=T) + geom_histogram(aes(y=..density..), alpha=
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

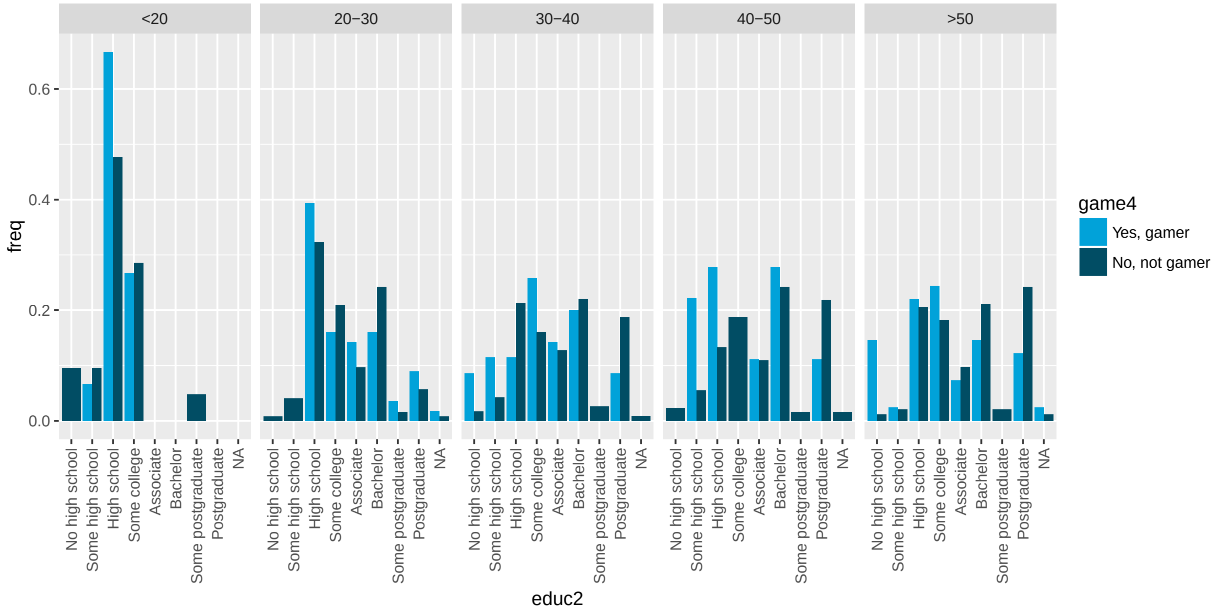


Unsurprisingly, gamers tend to be much younger than non-gamers.

Age likely has some impact on the above analyses, but it will much more likely come into play as we look at education levels, party affiliation and income. So before proceeding, we'll bin the sample by age groups to make it more manageable to deal with.

```
gamers = gamers %>%
  mutate(age_group = ifelse(age < 20, '<20',
                     ifelse(age < 30, '20-30',
                     ifelse(age < 40, '30-40',
                     ifelse(age < 50, '40-50',
                            '>50')))))
gamers$age_group = factor(gamers$age_group, levels = c('<20', '20-30', '30-40', '40-50', '>50'))
```

```
gamers %>%
  group_by(age_group, game4, educ2) %>%
  summarise(n = n()) %>%
  mutate(freq = n/ sum(n)) %>%
  ggplot(aes(x=educ2, y=freq, fill=game4)) +
  geom_bar(stat = 'identity', position = 'dodge') +
  facet_grid(. ~ age_group) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5)) +
  scale_fill_economist()
```



That's a bit of an eye chart, so let's try just plotting the difference in frequencies and at the same time demonstrate use of the spread function!

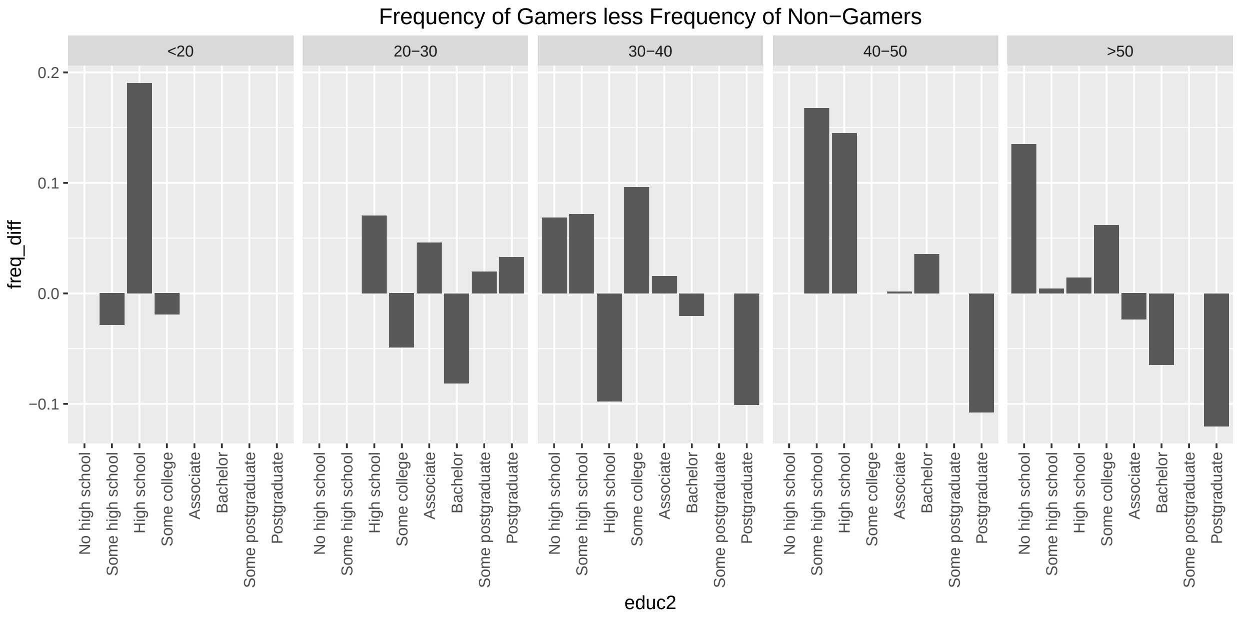
```
gamers_wide = gamers %>%
  group_by(age_group, game4, educ2) %>%
  summarise(n = n()) %>%
  mutate(freq = n/ sum(n))

gamers_wide = subset(gamers_wide, select=-c(n))  # drop n column
gamers_wide = gamers_wide %>%
  spread(game4, freq) %>%  # spread freq by gamer yes or no
  mutate(freq_diff = `Yes, gamer` - `No, not gamer`) %>%
  filter(!is.na(educ2))  # drop NAs

gamers_wide %>%
  ggplot(aes(x=educ2, y=freq_diff)) +
  geom_bar(stat = 'identity') +
  facet_grid(. ~ age_group) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5)) +
  ggtitle('Frequency of Gamers less Frequency of Non-Gamers')
```

```
## Warning: Removed 9 rows containing missing values (position_stack).
```

```
## Warning: Stacking not well defined when ymin != 0
```



Interestingly, I can't really 'eye-ball' any particular patterns in this data. For example, in the age groups >30, non-gamers are consistently more likely to have some amount of postgraduate experience than gamers. However, in the 20-30 age group, it's the reverse. There simply could be too few data points and thus too much noise to be able to read much from the data at this resolution level.

Let's see if that's true for income as well…

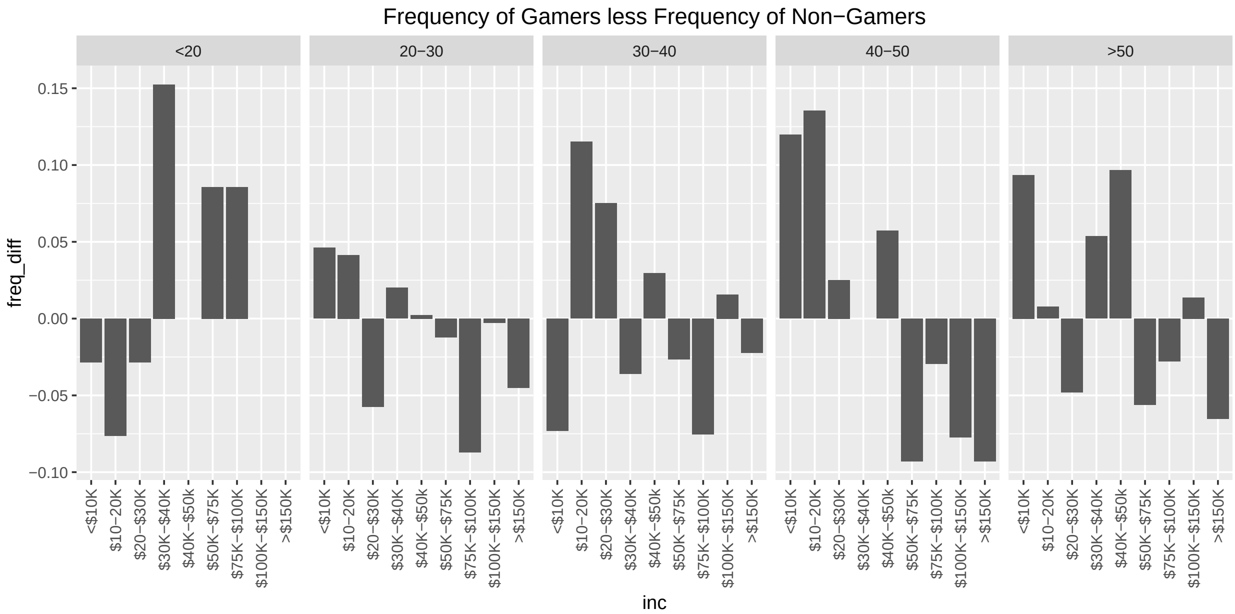
```
gamers_wide = gamers %>%
  group_by(age_group, game4, inc) %>%
  summarise(n = n()) %>%
  mutate(freq = n/ sum(n))

gamers_wide = subset(gamers_wide, select=-c(n))  # drop n column
gamers_wide = gamers_wide %>%
  spread(game4, freq) %>%  # spread freq by gamer yes or no
  mutate(freq_diff = `Yes, gamer` - `No, not gamer`) %>%
  filter(!is.na(inc))  # drop NAs

gamers_wide %>%
  ggplot(aes(x=inc, y=freq_diff)) +
  geom_bar(stat = 'identity') +
  facet_grid(. ~ age_group) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5)) +
  ggtitle('Frequency of Gamers less Frequency of Non-Gamers')
```

```
## Warning: Removed 4 rows containing missing values (position_stack).
```

```
## Warning: Stacking not well defined when ymin != 0
```

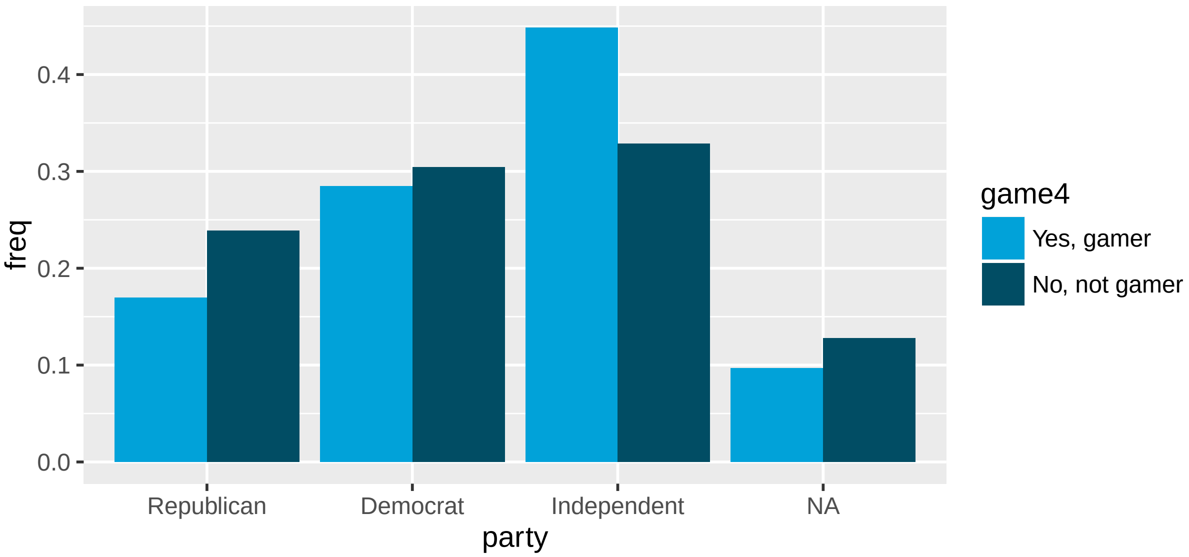


Here, we see a fairly consistent pattern of a lower frequency of gamers earning higher incomes in almost all by the youngest age groups, and conversely, gamers also disproportionately earn lower incomes.

Lastly, because it's an election year, let's have a look at party-affiliation for gamers and non-gamers…

```
plotGamers(gamers, 'party', 'game4', 'freq')
```

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



```
##
## [[2]]
##
##  3-sample test for equality of proportions without continuity
##  correction
```

```
##
## data:  table(gamers[, x], gamers[, group])
## X-squared = 8.1854, df = 2, p-value = 0.01669
## alternative hypothesis: two.sided
## sample estimates:
##    prop 1    prop 2    prop 3
## 0.1365854 0.1721612 0.2327044
```

The 3-sample proportion equality test p-value of 0.02 suggests that the proportion of gamers to non-gamers within each party group is statistically different. Surprisingly, the number of Independents within this sample is greater than than the number of Republicans and Democrats. It would be interesting to compare this with the population at large to see if there is some bias in our sample.