Mid-term Report Lab

Trang Nguyen, Dennis Liu

2/23/2017

INTRODUCTION

1. Data Scraping

Instead of downloading the csv from the website and then importing the file into R, we can stream live data using its url and the library **RCurl**.

```
library(RCurl)
```

```
## Loading required package: bitops
```

```
url <- "http://statgames.tietronix.com/tangrams/webreporter.php?game=tangrams&groupID=&winlose=both&ran
tangram <- read.csv(textConnection(readLines(url)), stringsAsFactors = F)
```

2. Data Cleaning

This is a messy dataset for a couple of reasons:
+ **TimeUsed**: There are many extreme outliers like 0 or 10000 seconds. Some users did not not finish the game and forgot to close the browser. Another issue with this variable is that time used is bounded above in some studies. If the study chooses restriction time request (60, 120, or 180 seconds), its time used will converge to the time limit and bias the results.
+ **Factors**: There are 3 self-input factors and Tangram set up them into 3 different columns. Therefore, we need to search through all 3 columns for a specific factor (i.e. Gender).

Step 1: Remove extreme outliers

```
tangram <- tangram[tangram$TimeUsed > 0 & tangram$TimeUsed < 3600,]
```

“‘ # Step 2: Use log transformation to reduce the variability of **TimeUsed** variable

```
tangram$TimeUsed_log <- log(tangram$TimeUsed)
```

Step 3: Consolidate 3 columns of factors and levels

As mentioned above, it is difficult to run analysis when there are 3 separate columns of factors and levels. Therefore, we run an algorithm over all 3 columns to identify some common solid factors and use them for further analysis.

```
factorlist <- as.data.frame(table(c(tangram$Factor1,tangram$Factor2, tangram$Factor3)))
factorlist <- factorlist[order(- factorlist$Freq),]
head(factorlist, 20)
```

```
##                  Var1   Freq
## 1                     125379
```

```
## 207            Gender   4174
## 461              STEM   1138
## 206            gender   1026
## 332             Major   1010
## 43                Age    598
## 209            GENDER    505
## 63            Athlete    391
## 308          LastName    318
## 67            Attempt    275
## 145              edad    253
## 3                   1    249
## 331             major    245
## 219            Genero    240
## 433          semestre    202
## 218            genero    194
## 529              Year    190
## 140 Do you own a pet?    179
## 42                age    177
## 146              Edad    158
```

After going through the list of factors, we decided to extract 4 factors: Gender, Athlete, STEM, and Age, using the algorithm below. We only include the code for Gender in this report (the code for other factors can be found in the **factor_filter.R** file) This algorithm can be broken down into 3 components:

```
### Factor to be filtered: GENDER
### Step 1: Create a key set: A set of criteria for regular expressions/ key patterns
gen = c("^gen","^sex")
gen_male = c("^m","^h")
gen_female = c("^f", "^mu")

### Step 2: Create a variable to identify whether a study includes GENDER as a factor or not
### Supporting functions
### ismatch
### @input: cond: a vector of key patterns (e.g. gen)
###         x, y, z: 3 columns to look for key
### @return: 0 if no match
###          index of the factor (1, 2, 3() if there is match
ismatch <- function(cond, x, y, z) {
  x <- grep(paste(cond, collapse = "|"), c(x, y, z), ignore.case = T, value = FALSE)
  return (ifelse(length(x), as.numeric(x), as.numeric(0)))
}

### Step 3: Create a variable for levels of genders if GENDER is a factor in that study
### level_gen_fun
### @input: x is the col index (given by factor_gender)
###         y is the row index
### @return: -1 if no gender factor indicated
###          -2 if gender factor indicated but level_gen key patterns not matched (require future manual
###           M/F
### @note: this function is specific to gender only. Needs to think about how to generalize it.
level_gen_fun <- function(x, y) {
  return (ifelse(x == 0, -1,
                 ifelse(grepl(paste(gen_female, collapse = "|"),tangram[y,2*x+3], ignore.case = T),
                        "F",
                        ifelse(grepl(paste(gen_male, collapse = "|"),tangram[y,2*x+3], ignore.case = T)
```

```
                                "M",-2))))
}
## cannot be used for data table
tangram <- as.data.table(tangram)
tangram[, factor_gender := ismatch(gen, Factor1, Factor2, Factor3), by = 1:nrow(tangram)]
```

```
##        Player.Alias        GroupName StudentAlias Factor1 Level1 Factor2
##     1:        48784
##     2:        48783
##     3:        48782
##     4:        48781
##     5:        48780 STATS250TANGTANG       JReall
##    ---
## 48758:            5             test       shonda
## 48759:            4             test       shonda
## 48760:            3
## 48761:            2       Greybeards    Dovahkiin
## 48762:            1
##        Level2 Factor3 Level3 NumShapes RequestedTime TimeUsed TimerDisplay
##     1:                                7             0  352.812            1
##     2:                                7             0   52.608            1
##     3:                                7             0   92.970            1
##     4:                                7             0  127.902            1
##     5:                                7             0   61.392            1
##    ---
## 48758:                                7             0   34.703            1
## 48759:                                7             0   33.843            1
## 48760:                                7             0   31.032            1
## 48761:                                7             0   86.446            1
## 48762:                                7             0  107.854            1
##        TimerHint NumClicks Won HintsEnabled HintsUsed      PuzzleName
##     1:         0       264   1            1         0    Laughing Man
##     2:         0        41   1            1         0        The Hook
##     3:         0        76   1            1         0   Andy's Puzzle
##     4:         0       122   1            1         0  The Hat Wearer
##     5:         0        51   1            0         0  A Simple Chair
##    ---
## 48758:         0        33   1            1         3    Laughing Man
## 48759:         0        33   1            1         0        The Hook
## 48760:         0        30   1            1         3        The Hook
## 48761:         0        46   1            1         1 Complex Hexagon
## 48762:         0        42   1            1         1   Andy's Puzzle
##                   Timestamp TimeUsed_log factor_gender
##     1: 2017-03-06 16:20:43     5.865935             0
##     2: 2017-03-06 16:12:39     3.962868             0
##     3: 2017-03-06 16:11:43     4.532277             0
##     4: 2017-03-06 16:10:05     4.851264             0
##     5: 2017-03-06 16:09:38     4.117280             0
##    ---
## 48758: 2012-07-16 13:19:10     3.546826             0
## 48759: 2012-07-16 13:18:08     3.521732             0
## 48760: 2012-07-16 13:17:18     3.435019             0
## 48761: 2012-07-13 14:04:40     4.459520             0
## 48762: 2012-07-13 14:02:27     4.680778             0
```

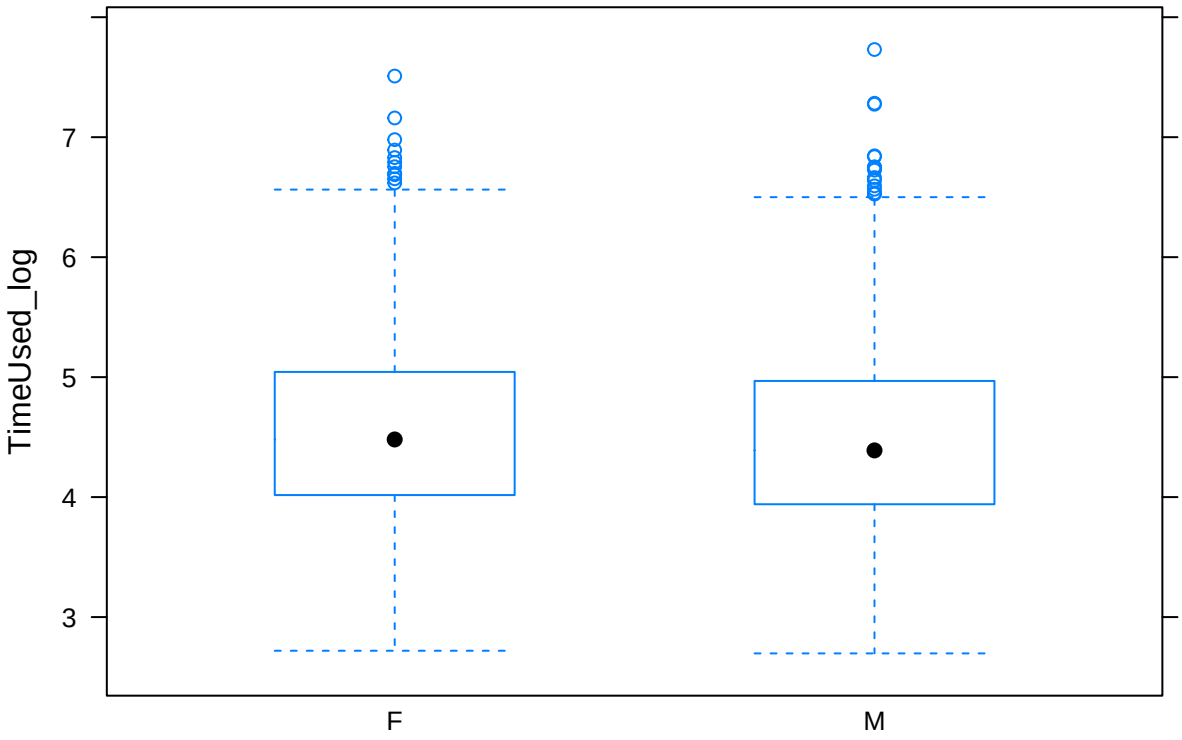
```
tangram <- as.data.frame(tangram)
tangram$level_gender <- mapply(level_gen_fun, tangram$factor_gender, 1:nrow(tangram), SIMPLIFY = T)
```

This algorithm has both pros and cons. Two main pros are its low time complexity and its generalizability across different types of factors. One cons is the divergence in using data table and data frame. More particularly, **ismatch** must be applied to datatable while **level_gen_fun** must be applied to a dataframe.

3. Data Exploration

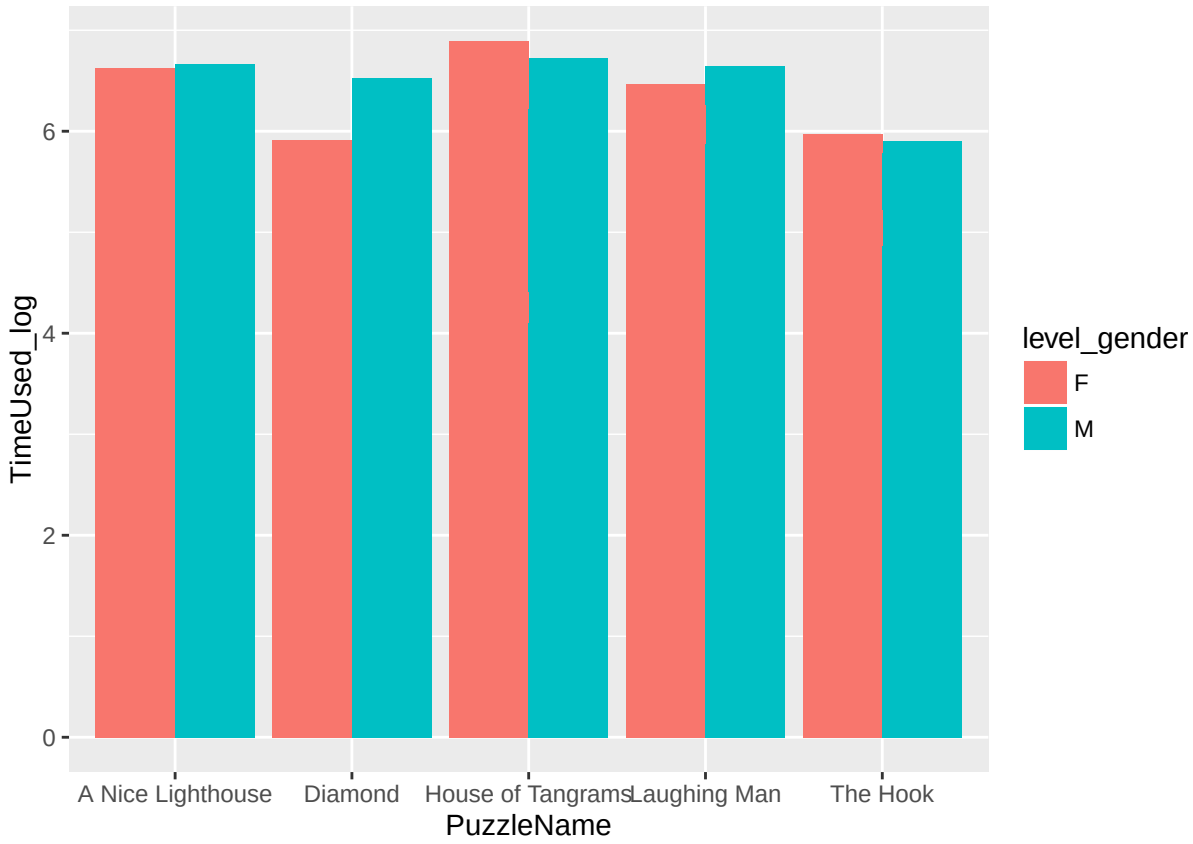
Approach 1: View the dataset as one study

```
bwplot( TimeUsed_log ~ level_gender, data = tangram[tangram$level_gender == "F"|tangram$level_gender ==
```



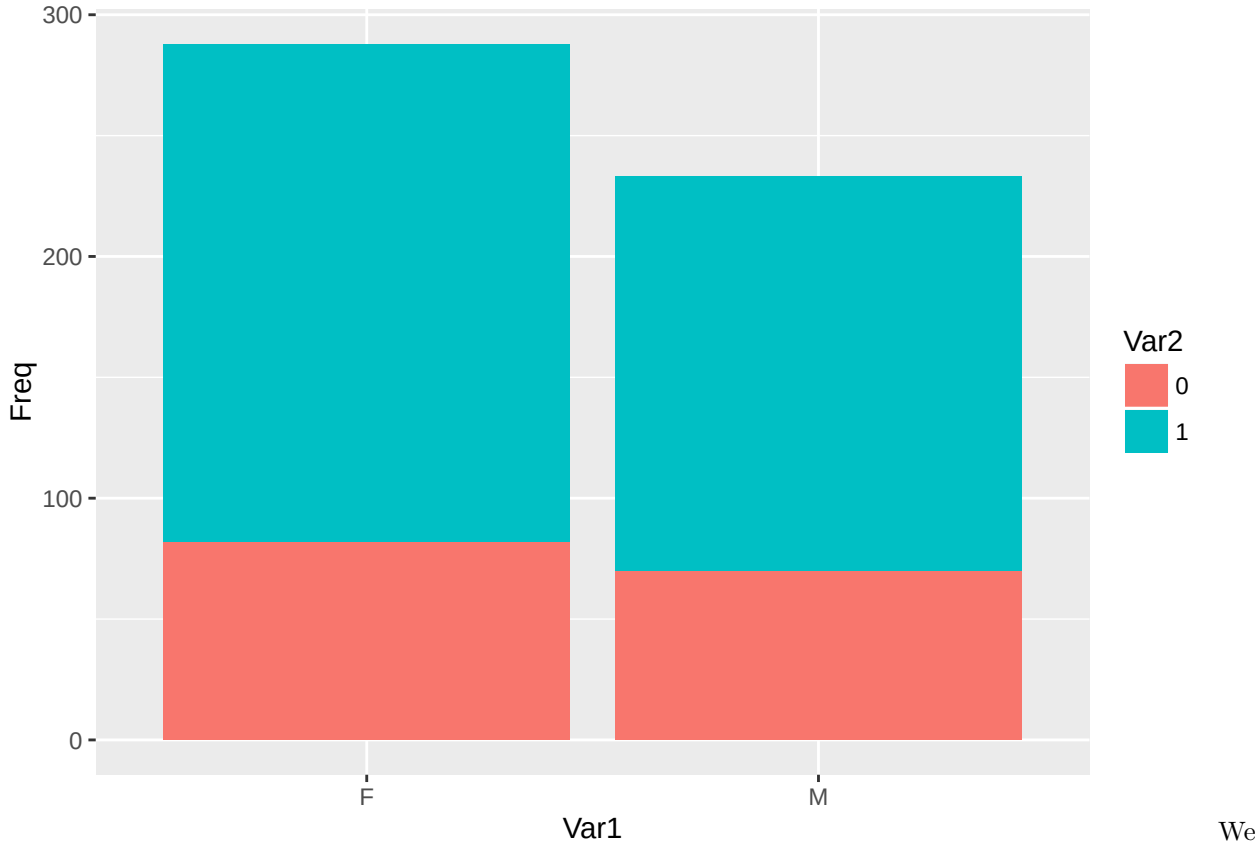
We are also interest in the gender difference within one game. In order to do that, we plot a double bar chart. With the y-axis as the average of Time_used and the X-axis as Puzzle Name.

```
gen_tangram <- tangram[tangram$level_gender == "F"|tangram$level_gender == "M",]
statsgame <- c("Laughing Man", "Piano", "The Hook", "Complex Hexagon", "Diamond", "House of Tangrams",
               "A Nice Lighthouse", "The Brain Buster", "A Simple Chair", "The Hat Wearer",
               "The Acrobat", "The Bird", "Crouching Cat", "The Goat", "The Six", "The G",
               "Andy's Puzzle", "Walking Person Puzzle", "A Medicine Jar", "Candle")
gen_tangram_statsgame <- gen_tangram[gen_tangram$PuzzleName %in% statsgame,]
gen_tangram_count <- count_(gen_tangram_statsgame,"PuzzleName")
gen_tangram_count <- gen_tangram_count[order(-gen_tangram_count$n),]
ggplot(gen_tangram_statsgame[gen_tangram_statsgame$PuzzleName %in% head(gen_tangram_count$PuzzleName,5)
  geom_bar(position = "dodge", stat = "identity")
```



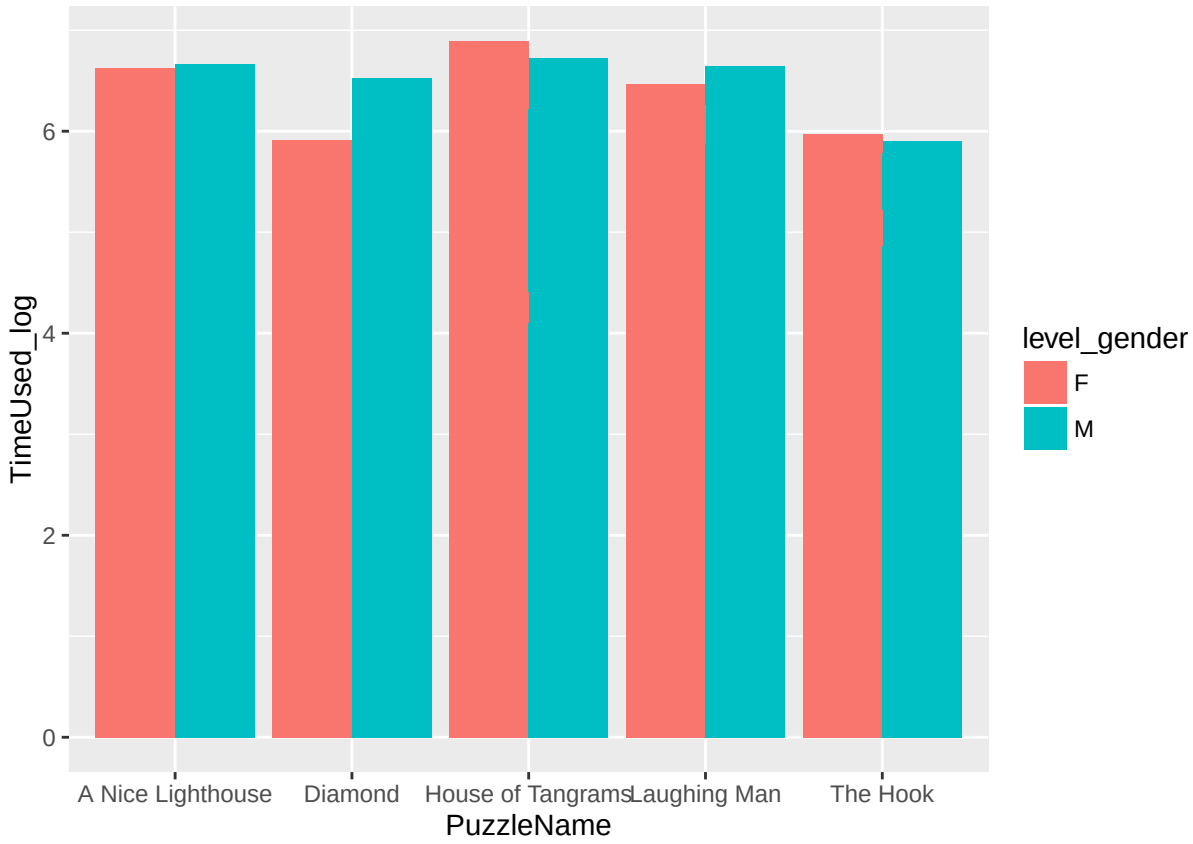
We noticed that there are some samples request a limited time. If they request a limited time, there is a possibility such that they will fail the game. From that, we have the plot as follows,

```
gen_won_res_count <-  as.data.frame(table(gen_tangram[gen_tangram$RequestedTime != 0,]$level_gender,gen
ggplot(gen_won_res_count, aes(x=Var1, y=Freq,fill=Var2)) + geom_bar(stat="identity")
```



We are also interest in the gender difference within one game. In order to do that, we plot a double bar chart. With the y-axis as the average of Time_used and the X-axis as Puzzle Name.

```
gen_tangram_statsgame <- gen_tangram[gen_tangram$PuzzleName %in% statsgame,]
gen_tangram_count <- count_(gen_tangram_statsgame,"PuzzleName")
gen_tangram_count <- gen_tangram_count[order(-gen_tangram_count$n),]
ggplot(gen_tangram_statsgame[gen_tangram_statsgame$PuzzleName %in% head(gen_tangram_count$PuzzleName,5)
  geom_bar(position = "dodge", stat = "identity")
```

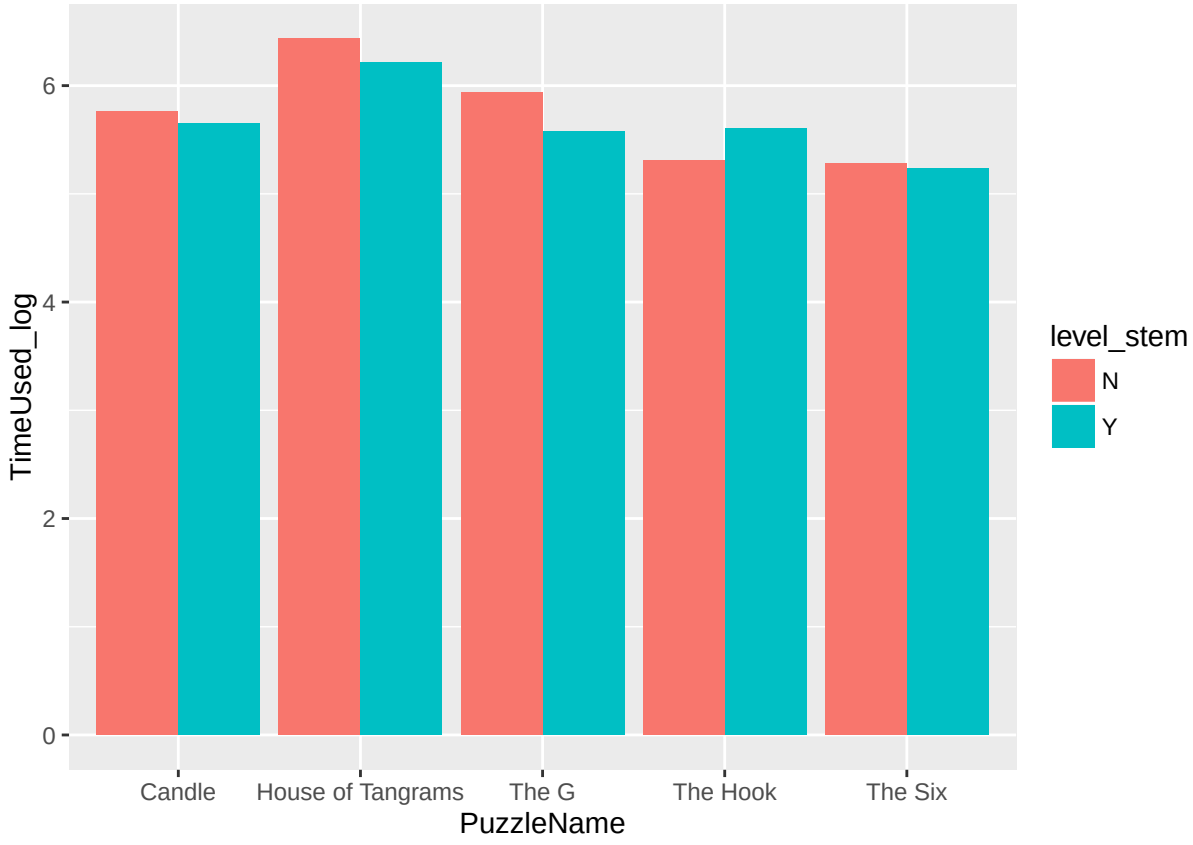


Another factor we are having is STEM major.

```
### Factor to be filtered: STEM
### Step 1: Create a key set: A set of criteria for regular expressions/ key patterns
stem <- c("stem")
stem_Y <- c("^y")
stem_N <- c("^n")
### Step 2: Create a variable to identify whether a study includes GENDER as a factor or not

tangram$factor_STEM <- mapply(ismatch, stem, tangram$Factor1, tangram$Factor2, tangram$Factor3, SIMPLIF
### Step 3: Create a variable for levels of STEM if STEM is a factor in that study
### level_stem_fun
### @input: x is the col index (given by factor_gender)
###         y is the row index
### @return: -1 if no gender factor indicated
###          -2 if gender factor indicated but level_gen key patterns not matched (require future manua

###           Y/N
### @note: this function is specific to STEM only. Needs to think about how to generalize it.
level_stem_fun <- function(x, y) {
  return (ifelse(x == 0, -1,
                 ifelse(grepl(paste(stem_Y, collapse = "|"),tangram[y,2*x+3], ignore.case = T),
                        "Y",
                        ifelse(grepl(paste(stem_N, collapse = "|"),tangram[y,2*x+3], ignore.case = T),
                               "N",-2))))
}

tangram$level_stem <- mapply(level_stem_fun, tangram$factor_STEM, 1:nrow(tangram), SIMPLIFY = TRUE)
```

```
stem_tangram <- tangram[tangram$level_stem == "Y"|tangram$level_stem == "N",]
```

```
stem_tangram_statsgame <- stem_tangram[stem_tangram$PuzzleName %in% statsgame,]
stem_tangram_count <- count_(stem_tangram_statsgame,"PuzzleName")
stem_tangram_count <- stem_tangram_count[order(-stem_tangram_count$n),]
ggplot(stem_tangram_statsgame[stem_tangram_statsgame$PuzzleName %in% head(stem_tangram_count$PuzzleName
  geom_bar(position = "dodge", stat = "identity")
```



Approach 2: View the dataset as a cumulation of different studies

This approach views the whole dataset as a collection of different samples. We tested this approach for the question: *"Does gender influence the time used?"* ran through all samples and calculated p-value for each sample to see the distribution of p-value in the whole dataset. The variability of p-value in this distribution aims at demonstrating to statistics students how unreliable p-value can be sometimes. Some people argue that that increasing sample size is the only thing we need to in order to have a reliable p-value. However, in the scatterplot of sample size and p-value, we can see that there is not a clear correlation. There are other factors in design of experiments researchers need to pay attention to in order to guarantee the reliability of p-value.

```
tangram_tested <- as.data.table(tangram)
tangram_tested$level_gender[tangram_tested$level_gender == -2] <- "Others"
tangram_tested$GroupName <- tolower(tangram_tested$GroupName) ### ASSUMPTION: case-insenstive for group
# filtering out null names and groups with small sample size (< 5)
tangram_tested <- tangram_tested[tangram_tested$GroupName != "",]
```

```r
# Only include 2 levels of gender: Female and Male
tangram_pvalue <- tangram_tested[tangram_tested$level_gender %in% c("F", "M"),]
tangram_pvalue[, `:=`( SampleSize = .N) , by = GroupName]
```

```
##       Player.Alias GroupName StudentAlias Factor1 Level1   Factor2 Level2
##    1:        48080    2017t1      PRUEBA1  GENERO      H PROFESION     NI
##    2:        48079    2017t1      ENRIQUE  GENERO      H PROFESION      I
##    3:        47952   bubbles      cikocis  gender   male
##    4:        47949   bubbles      cikocis  gender   male
##    5:        47947   bubbles      cikocis  gender   male
##   ---
## 6410:           86    mt6020       xpxp12  gender      m       exp     no
## 6411:           84    mt6020      1984mrf  gender      M  Priorexp      N
## 6412:           83    mt6020      1984mrf  gender      M  Priorexp      N
## 6413:           81    mt6020      dfreawy  gender      m       exp     no
## 6414:           10      test   Iamsocoool  Gender   Male   Athlete     No
##       Factor3 Level3 NumShapes RequestedTime TimeUsed TimerDisplay
##    1:                        7             0   33.601            1
##    2:                        7             0   37.686            1
##    3:                        7             0   37.665            1
##    4:                        7             0   37.842            1
##    5:                        7             0   34.322            1
##   ---
## 6410:                        7             0   64.670            1
## 6411:                        7             0  126.122            0
## 6412:                        7             0  267.388            0
## 6413:                        7             0  107.142            1
## 6414:                        7             0   60.338            0
##       TimerHint NumClicks Won HintsEnabled HintsUsed            PuzzleName
##    1:         0        33   1            1         0              The Hook
##    2:         0        25   1            1         0              The Hook
##    3:         0        37   1            0         0               The Six
##    4:         0        56   1            0         0                 The G
##    5:         0        35   1            0         0         Andy's Puzzle
##   ---
## 6410:         0        25   1            1         0               Candle
## 6411:         0        50   1            1         0              Diamond
## 6412:         0       124   1            1         0          Laughing Man
## 6413:         0        40   1            1         6 Walking Person Puzzle
## 6414:         0        35   1            0         0     House of Tangrams
##                 Timestamp TimeUsed_log factor_gender level_gender
##    1: 2017-03-03 14:58:01     3.514556             1            M
##    2: 2017-03-03 14:56:49     3.629289             1            M
##    3: 2017-03-03 07:30:34     3.628731             1            M
##    4: 2017-03-03 07:29:20     3.633420             1            M
##    5: 2017-03-03 07:26:47     3.535787             1            M
##   ---
## 6410: 2012-10-21 23:49:49     4.169297             1            M
## 6411: 2012-10-21 21:41:28     4.837250             1            M
## 6412: 2012-10-21 21:39:07     5.588701             1            M
## 6413: 2012-10-13 12:50:25     4.674155             1            M
## 6414: 2012-07-24 08:14:39     4.099962             1            M
##       factor_STEM level_stem SampleSize
```

```
##    1:           0         -1          2
##    2:           0         -1          2
##    3:           0         -1         10
##    4:           0         -1         10
##    5:           0         -1         10
##   ---
## 6410:           0         -1          5
## 6411:           0         -1          5
## 6412:           0         -1          5
## 6413:           0         -1          5
## 6414:           0         -1         70
```

```
tangram_pvalue <- tangram_pvalue[tangram_pvalue$SampleSize >= 5,]
tangram_pvalue[, `:=`(log_timeused_mean = mean(TimeUsed_log), timeused_mean = mean(TimeUsed)), by = c("
```

```
##       Player.Alias GroupName StudentAlias Factor1 Level1  Factor2 Level2
##    1:        47952   bubbles      cikocis  gender   male
##    2:        47949   bubbles      cikocis  gender   male
##    3:        47947   bubbles      cikocis  gender   male
##    4:        47945   bubbles      cikocis  gender   male
##    5:        47942   bubbles      cikocis  gender   male
##   ---
## 6142:           86    mt6020       xpxp12  gender      m      exp     no
## 6143:           84    mt6020      1984mrf  gender      M Priorexp      N
## 6144:           83    mt6020      1984mrf  gender      M Priorexp      N
## 6145:           81    mt6020      dfreawy  gender      m      exp     no
## 6146:           10      test   Iamsocoool  Gender   Male  Athlete     No
##       Factor3 Level3 NumShapes RequestedTime TimeUsed TimerDisplay
##    1:                       7             0   37.665            1
##    2:                       7             0   37.842            1
##    3:                       7             0   34.322            1
##    4:                       7             0   36.211            1
##    5:                       7             0  143.260            1
##   ---
## 6142:                       7             0   64.670            1
## 6143:                       7             0  126.122            0
## 6144:                       7             0  267.388            0
## 6145:                       7             0  107.142            1
## 6146:                       7             0   60.338            0
##       TimerHint NumClicks Won HintsEnabled HintsUsed           PuzzleName
##    1:         0        37   1            0         0              The Six
##    2:         0        56   1            0         0                The G
##    3:         0        35   1            0         0        Andy's Puzzle
##    4:         0        38   1            0         0       A Medicine Jar
##    5:         0       141   1            0         0               Candle
##   ---
## 6142:         0        25   1            1         0               Candle
## 6143:         0        50   1            1         0              Diamond
## 6144:         0       124   1            1         0         Laughing Man
## 6145:         0        40   1            1         6 Walking Person Puzzle
## 6146:         0        35   1            0         0     House of Tangrams
##                  Timestamp TimeUsed_log factor_gender level_gender
##    1: 2017-03-03 07:30:34     3.628731             1            M
##    2: 2017-03-03 07:29:20     3.633420             1            M
##    3: 2017-03-03 07:26:47     3.535787             1            M
```

```
##    4: 2017-03-03 07:24:29     3.589363             1            M
##    5: 2017-03-03 07:23:43     4.964661             1            M
##   ---
## 6142: 2012-10-21 23:49:49     4.169297             1            M
## 6143: 2012-10-21 21:41:28     4.837250             1            M
## 6144: 2012-10-21 21:39:07     5.588701             1            M
## 6145: 2012-10-13 12:50:25     4.674155             1            M
## 6146: 2012-07-24 08:14:39     4.099962             1            M
##       factor_STEM level_stem SampleSize log_timeused_mean timeused_mean
##    1:           0         -1         10          4.001152      60.55480
##    2:           0         -1         10          4.001152      60.55480
##    3:           0         -1         10          4.001152      60.55480
##    4:           0         -1         10          4.001152      60.55480
##    5:           0         -1         10          4.001152      60.55480
##   ---
## 6142:           0         -1          5          4.633952     122.94840
## 6143:           0         -1          5          4.633952     122.94840
## 6144:           0         -1          5          4.633952     122.94840
## 6145:           0         -1          5          4.633952     122.94840
## 6146:           0         -1         70          4.197110      80.63254
```

```
tangram_pvalue[, `:=`(n = .N, sd = sd(TimeUsed_log)), by = c("GroupName", "level_gender")]
```

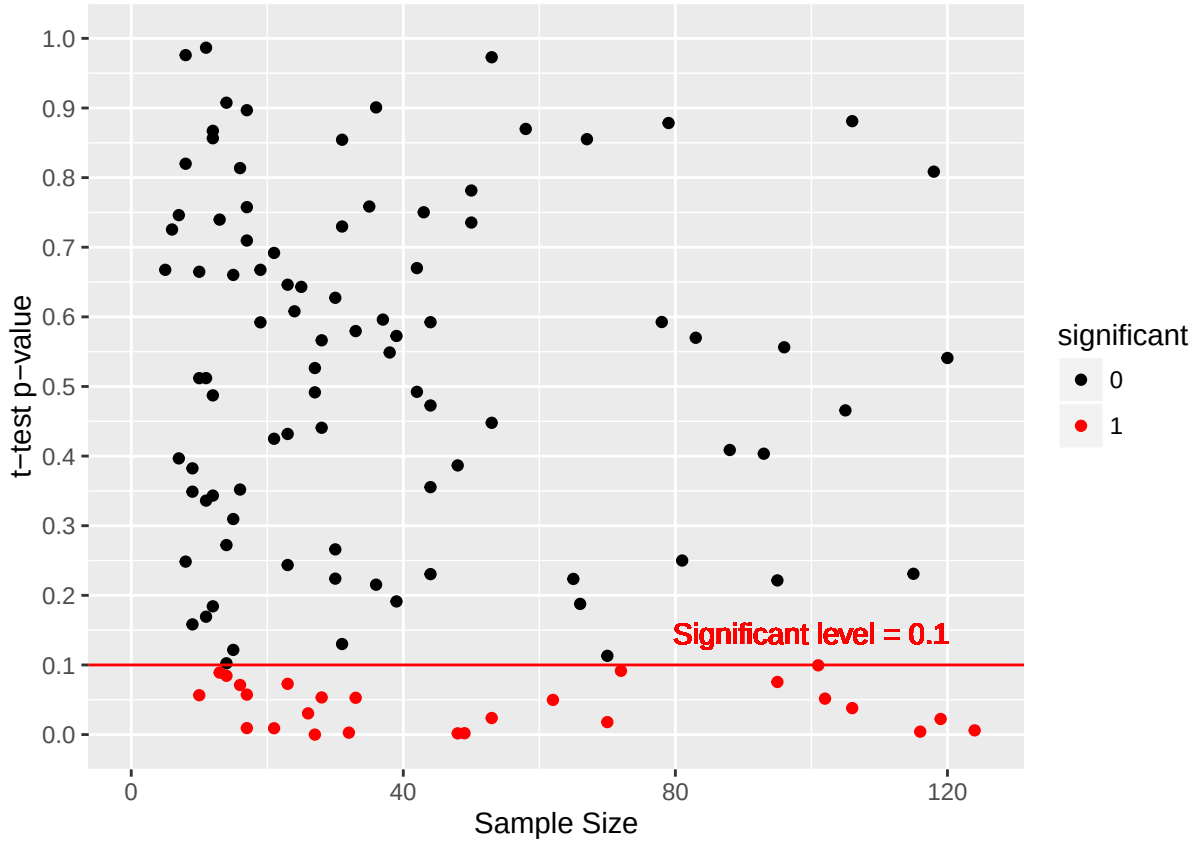
```
##       Player.Alias GroupName StudentAlias Factor1 Level1  Factor2 Level2
##    1:        47952   bubbles      cikocis  gender   male
##    2:        47949   bubbles      cikocis  gender   male
##    3:        47947   bubbles      cikocis  gender   male
##    4:        47945   bubbles      cikocis  gender   male
##    5:        47942   bubbles      cikocis  gender   male
##   ---
## 6142:           86    mt6020       xpxp12  gender      m      exp     no
## 6143:           84    mt6020      1984mrf  gender      M Priorexp      N
## 6144:           83    mt6020      1984mrf  gender      M Priorexp      N
## 6145:           81    mt6020      dfreawy  gender      m      exp     no
## 6146:           10      test   Iamsocoool  Gender   Male  Athlete     No
##       Factor3 Level3 NumShapes RequestedTime TimeUsed TimerDisplay
##    1:                        7             0   37.665            1
##    2:                        7             0   37.842            1
##    3:                        7             0   34.322            1
##    4:                        7             0   36.211            1
##    5:                        7             0  143.260            1
##   ---
## 6142:                        7             0   64.670            1
## 6143:                        7             0  126.122            0
## 6144:                        7             0  267.388            0
## 6145:                        7             0  107.142            1
## 6146:                        7             0   60.338            0
##       TimerHint NumClicks Won HintsEnabled HintsUsed           PuzzleName
##    1:         0        37   1            0         0              The Six
##    2:         0        56   1            0         0                The G
##    3:         0        35   1            0         0        Andy's Puzzle
##    4:         0        38   1            0         0       A Medicine Jar
##    5:         0       141   1            0         0               Candle
##   ---
## 6142:         0        25   1            1         0               Candle
```

```
## 6143:         0        50   1            1         0               Diamond
## 6144:         0       124   1            1         0         Laughing Man
## 6145:         0        40   1            1         6 Walking Person Puzzle
## 6146:         0        35   1            0         0     House of Tangrams
##                 Timestamp TimeUsed_log factor_gender level_gender
##    1: 2017-03-03 07:30:34     3.628731             1            M
##    2: 2017-03-03 07:29:20     3.633420             1            M
##    3: 2017-03-03 07:26:47     3.535787             1            M
##    4: 2017-03-03 07:24:29     3.589363             1            M
##    5: 2017-03-03 07:23:43     4.964661             1            M
##   ---
## 6142: 2012-10-21 23:49:49     4.169297             1            M
## 6143: 2012-10-21 21:41:28     4.837250             1            M
## 6144: 2012-10-21 21:39:07     5.588701             1            M
## 6145: 2012-10-13 12:50:25     4.674155             1            M
## 6146: 2012-07-24 08:14:39     4.099962             1            M
##       factor_STEM level_stem SampleSize log_timeused_mean timeused_mean  n
##    1:           0         -1         10          4.001152      60.55480 10
##    2:           0         -1         10          4.001152      60.55480 10
##    3:           0         -1         10          4.001152      60.55480 10
##    4:           0         -1         10          4.001152      60.55480 10
##    5:           0         -1         10          4.001152      60.55480 10
##   ---
## 6142:           0         -1          5          4.633952     122.94840  5
## 6143:           0         -1          5          4.633952     122.94840  5
## 6144:           0         -1          5          4.633952     122.94840  5
## 6145:           0         -1          5          4.633952     122.94840  5
## 6146:           0         -1         70          4.197110      80.63254 50
##              sd
##    1: 0.4510175
##    2: 0.4510175
##    3: 0.4510175
##    4: 0.4510175
##    5: 0.4510175
##   ---
## 6142: 0.6535594
## 6143: 0.6535594
## 6144: 0.6535594
## 6145: 0.6535594
## 6146: 0.6128304
```

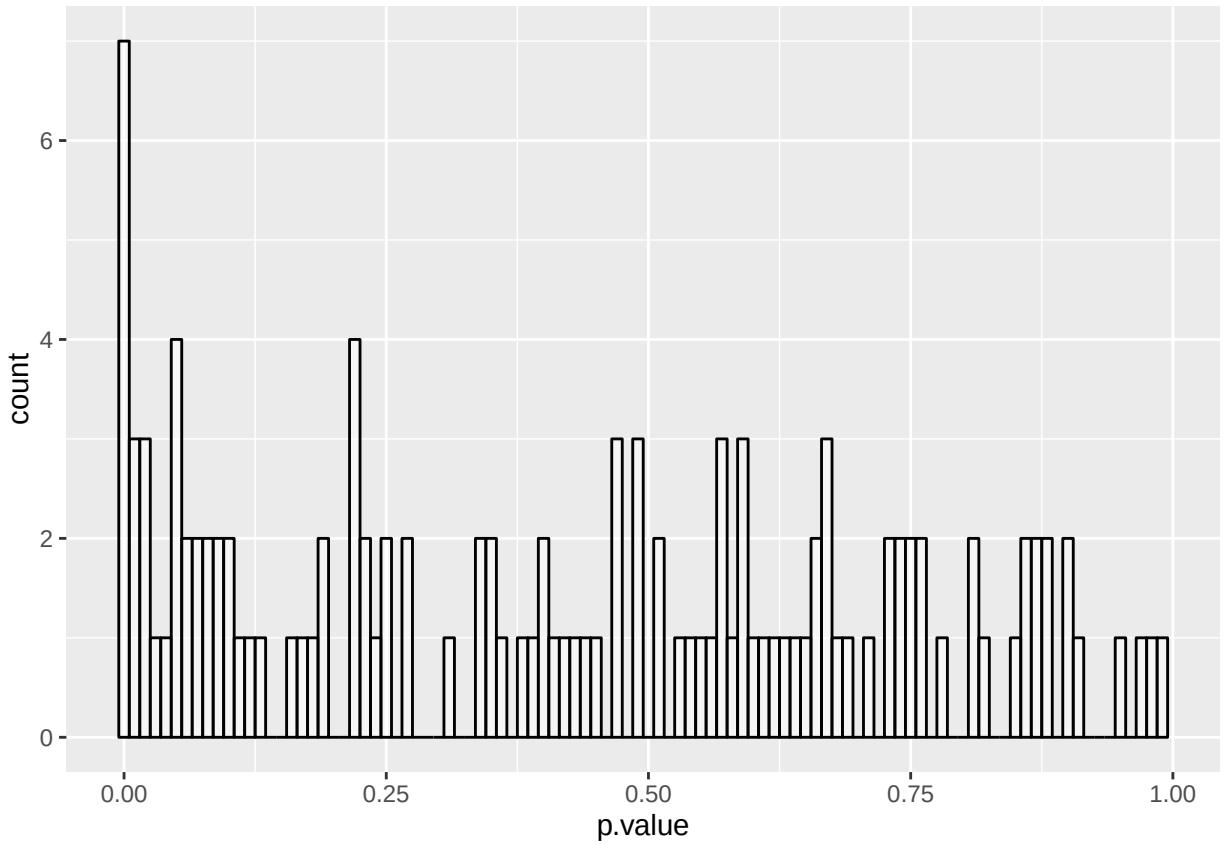
```r
ttestFun <- function(dat) {
  if (sum(dat$level_gender == "F") > 1 && sum(dat$level_gender == "M") > 1) {
    the_fit <- t.test(TimeUsed_log ~ level_gender, data = dat)
    #setNames(the_fit$p.value, "p.value")}
    c("p.value" = the_fit$p.value, "samplesize" = mean(dat$SampleSize))}
  else {
    c("p.value" = -1, "samplesize" = mean(dat$SampleSize))
  }
}
# 1. Scatterplot of pvalue vs. sample size

alpha <- 0.1
gender_pval_dist <- ddply(tangram_pvalue, ~ GroupName, ttestFun)
```

```
gender_pval_dist$significant <-(gender_pval_dist$p.value < alpha)*1
gender_pval_dist$significant <- as.factor(gender_pval_dist$significant)
ggplot(gender_pval_dist[gender_pval_dist$p.value != -1,], aes(x = samplesize, y = p.value)) +
  geom_point(aes(colour = significant)) +
  scale_color_manual(values=c("black", "red")) +
  geom_hline(aes(yintercept = alpha), color = "red") +
  geom_text(aes(100,0.1,label = "Significant level = 0.1", vjust = -1), color = "red") +
  scale_x_continuous(name="Sample Size", limits=c(0, 125)) +
  scale_y_continuous(name="t-test p-value", limits=c(0, 1), breaks = seq(0,1, by = 0.1))
```

```
## Warning: Removed 6 rows containing missing values (geom_point).
```




```
# 2. Histogram of p-value
ggplot(gender_pval_dist[gender_pval_dist$p.value != -1,], aes(p.value)) +
  geom_histogram(binwidth = 0.01, position = "identity", alpha = 0.5, colour = "black", fill = "white")
```




```
# 3. Overlaid histogram for means (Female vs Male)
gender_samplingdis <- tangram_tested[tangram_tested$level_gender != -1,]
gender_samplingdis[,`:=`(SampleSize = .N), by = GroupName]
```

```
##       Player.Alias GroupName StudentAlias Factor1 Level1   Factor2 Level2
##    1:        48080    2017t1      PRUEBA1  GENERO      H PROFESION     NI
##    2:        48079    2017t1      ENRIQUE  GENERO      H PROFESION      I
##    3:        47952   bubbles      cikocis  gender   male
##    4:        47949   bubbles      cikocis  gender   male
##    5:        47947   bubbles      cikocis  gender   male
##   ---
## 6638:           86    mt6020       xpxp12  gender      m       exp     no
## 6639:           84    mt6020      1984mrf  gender      M  Priorexp      N
## 6640:           83    mt6020      1984mrf  gender      M  Priorexp      N
## 6641:           81    mt6020      dfreawy  gender      m       exp     no
## 6642:           10      test   Iamsocoool  Gender   Male   Athlete     No
##       Factor3 Level3 NumShapes RequestedTime TimeUsed TimerDisplay
##    1:                        7             0   33.601            1
##    2:                        7             0   37.686            1
##    3:                        7             0   37.665            1
##    4:                        7             0   37.842            1
##    5:                        7             0   34.322            1
##   ---
## 6638:                        7             0   64.670            1
## 6639:                        7             0  126.122            0
## 6640:                        7             0  267.388            0
## 6641:                        7             0  107.142            1
```

```
## 6642:                        7            0   60.338           0
##       TimerHint NumClicks Won HintsEnabled HintsUsed           PuzzleName
##    1:         0        33   1            1         0             The Hook
##    2:         0        25   1            1         0             The Hook
##    3:         0        37   1            0         0              The Six
##    4:         0        56   1            0         0                The G
##    5:         0        35   1            0         0        Andy's Puzzle
##   ---
## 6638:         0        25   1            1         0               Candle
## 6639:         0        50   1            1         0              Diamond
## 6640:         0       124   1            1         0         Laughing Man
## 6641:         0        40   1            1         6 Walking Person Puzzle
## 6642:         0        35   1            0         0     House of Tangrams
##                 Timestamp TimeUsed_log factor_gender level_gender
##    1: 2017-03-03 14:58:01     3.514556             1            M
##    2: 2017-03-03 14:56:49     3.629289             1            M
##    3: 2017-03-03 07:30:34     3.628731             1            M
##    4: 2017-03-03 07:29:20     3.633420             1            M
##    5: 2017-03-03 07:26:47     3.535787             1            M
##   ---
## 6638: 2012-10-21 23:49:49     4.169297             1            M
## 6639: 2012-10-21 21:41:28     4.837250             1            M
## 6640: 2012-10-21 21:39:07     5.588701             1            M
## 6641: 2012-10-13 12:50:25     4.674155             1            M
## 6642: 2012-07-24 08:14:39     4.099962             1            M
##       factor_STEM level_stem SampleSize
##    1:           0         -1          2
##    2:           0         -1          2
##    3:           0         -1         10
##    4:           0         -1         10
##    5:           0         -1         10
##   ---
## 6638:           0         -1          5
## 6639:           0         -1          5
## 6640:           0         -1          5
## 6641:           0         -1          5
## 6642:           0         -1         70
```

```
gender_samplingdis[, `:=`(log_timeused_mean = mean(TimeUsed_log), timeused_mean = mean(TimeUsed)), by =
```

```
##       Player.Alias GroupName StudentAlias Factor1 Level1   Factor2 Level2
##    1:        48080    2017t1      PRUEBA1  GENERO      H PROFESION     NI
##    2:        48079    2017t1      ENRIQUE  GENERO      H PROFESION      I
##    3:        47952   bubbles      cikocis  gender   male
##    4:        47949   bubbles      cikocis  gender   male
##    5:        47947   bubbles      cikocis  gender   male
##   ---
## 6638:           86    mt6020       xpxp12  gender      m       exp     no
## 6639:           84    mt6020      1984mrf  gender      M  Priorexp      N
## 6640:           83    mt6020      1984mrf  gender      M  Priorexp      N
## 6641:           81    mt6020      dfreawy  gender      m       exp     no
## 6642:           10      test   Iamsocoool  Gender   Male   Athlete     No
##       Factor3 Level3 NumShapes RequestedTime TimeUsed TimerDisplay
##    1:                        7             0   33.601            1
##    2:                        7             0   37.686            1
```

```
##    3:                        7             0   37.665           1
##    4:                        7             0   37.842           1
##    5:                        7             0   34.322           1
##   ---
## 6638:                        7             0   64.670           1
## 6639:                        7             0  126.122           0
## 6640:                        7             0  267.388           0
## 6641:                        7             0  107.142           1
## 6642:                        7             0   60.338           0
##       TimerHint NumClicks Won HintsEnabled HintsUsed            PuzzleName
##    1:         0        33   1            1         0              The Hook
##    2:         0        25   1            1         0              The Hook
##    3:         0        37   1            0         0               The Six
##    4:         0        56   1            0         0                 The G
##    5:         0        35   1            0         0         Andy's Puzzle
##   ---
## 6638:         0        25   1            1         0                Candle
## 6639:         0        50   1            1         0               Diamond
## 6640:         0       124   1            1         0          Laughing Man
## 6641:         0        40   1            1         6 Walking Person Puzzle
## 6642:         0        35   1            0         0     House of Tangrams
##                  Timestamp TimeUsed_log factor_gender level_gender
##    1: 2017-03-03 14:58:01     3.514556             1            M
##    2: 2017-03-03 14:56:49     3.629289             1            M
##    3: 2017-03-03 07:30:34     3.628731             1            M
##    4: 2017-03-03 07:29:20     3.633420             1            M
##    5: 2017-03-03 07:26:47     3.535787             1            M
##   ---
## 6638: 2012-10-21 23:49:49     4.169297             1            M
## 6639: 2012-10-21 21:41:28     4.837250             1            M
## 6640: 2012-10-21 21:39:07     5.588701             1            M
## 6641: 2012-10-13 12:50:25     4.674155             1            M
## 6642: 2012-07-24 08:14:39     4.099962             1            M
##       factor_STEM level_stem SampleSize log_timeused_mean timeused_mean
##    1:           0         -1          2          3.571922      35.64350
##    2:           0         -1          2          3.571922      35.64350
##    3:           0         -1         10          4.001152      60.55480
##    4:           0         -1         10          4.001152      60.55480
##    5:           0         -1         10          4.001152      60.55480
##   ---
## 6638:           0         -1          5          4.633952     122.94840
## 6639:           0         -1          5          4.633952     122.94840
## 6640:           0         -1          5          4.633952     122.94840
## 6641:           0         -1          5          4.633952     122.94840
## 6642:           0         -1         70          4.197110      80.63254
```

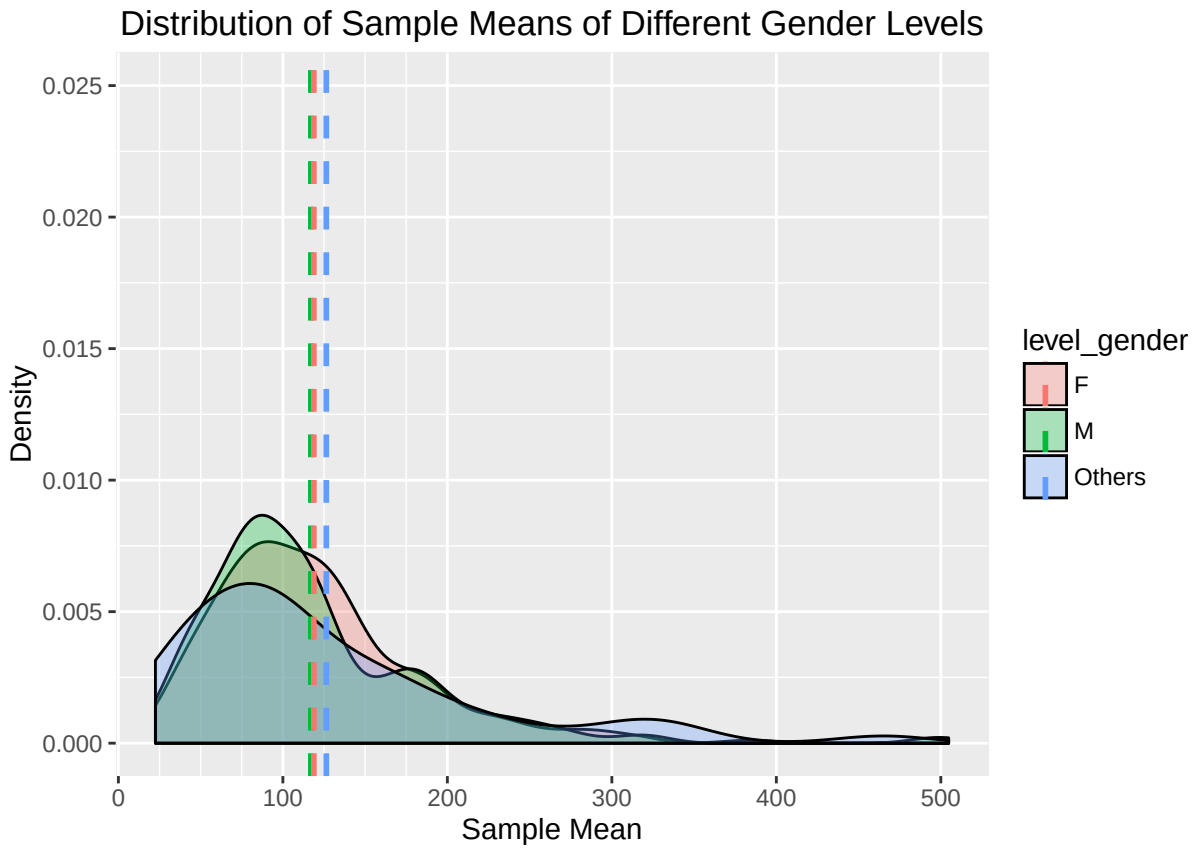
```
gender_samplingdis[, `:=`(n = .N, sd = sd(TimeUsed_log)), by = c("GroupName", "level_gender")]
```

```
##       Player.Alias GroupName StudentAlias Factor1 Level1   Factor2 Level2
##    1:        48080    2017t1      PRUEBA1  GENERO      H PROFESION     NI
##    2:        48079    2017t1      ENRIQUE  GENERO      H PROFESION      I
##    3:        47952   bubbles      cikocis  gender   male
##    4:        47949   bubbles      cikocis  gender   male
##    5:        47947   bubbles      cikocis  gender   male
##   ---
```

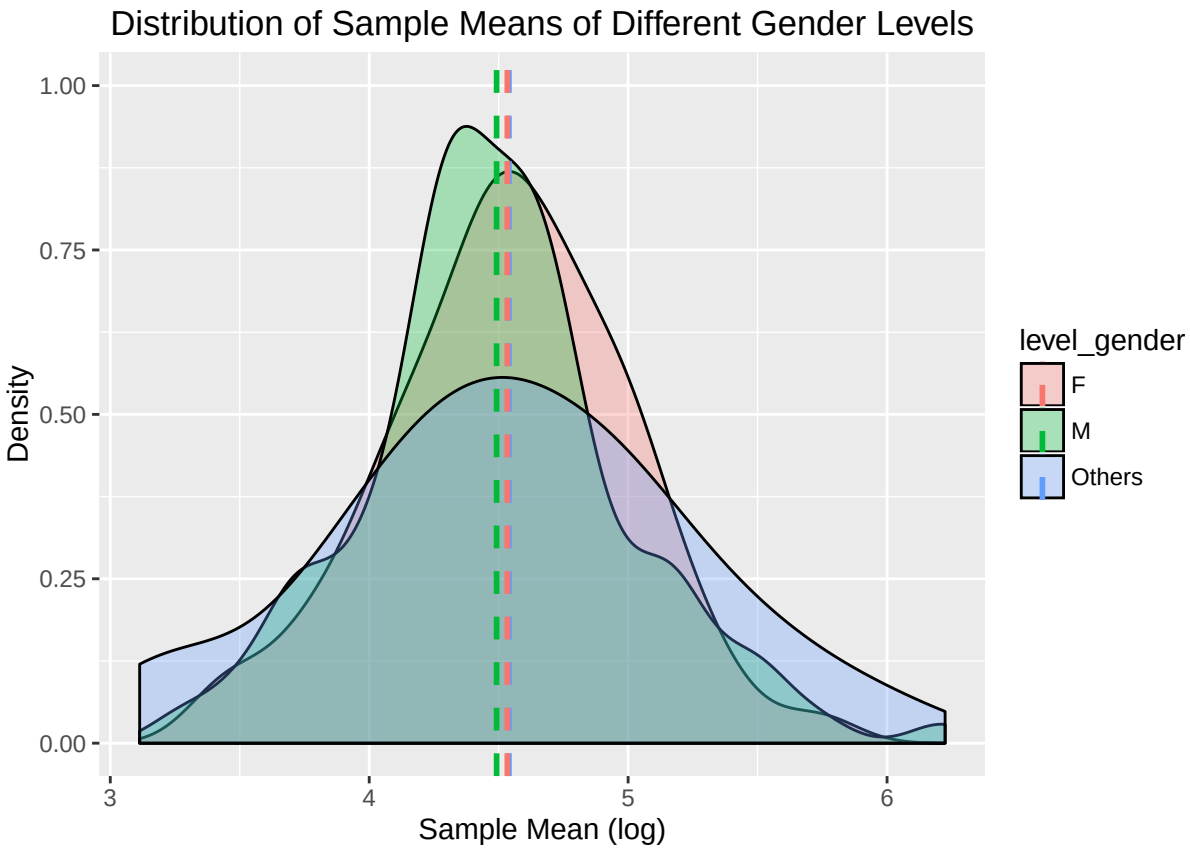
```
## 6638:           86    mt6020       xpxp12  gender      m       exp     no
## 6639:           84    mt6020      1984mrf  gender      M  Priorexp      N
## 6640:           83    mt6020      1984mrf  gender      M  Priorexp      N
## 6641:           81    mt6020      dfreawy  gender      m       exp     no
## 6642:           10      test   Iamsocoool  Gender   Male   Athlete     No
##       Factor3 Level3 NumShapes RequestedTime TimeUsed TimerDisplay
##    1:                        7             0   33.601            1
##    2:                        7             0   37.686            1
##    3:                        7             0   37.665            1
##    4:                        7             0   37.842            1
##    5:                        7             0   34.322            1
##   ---
## 6638:                        7             0   64.670            1
## 6639:                        7             0  126.122            0
## 6640:                        7             0  267.388            0
## 6641:                        7             0  107.142            1
## 6642:                        7             0   60.338            0
##       TimerHint NumClicks Won HintsEnabled HintsUsed            PuzzleName
##    1:         0        33   1            1         0              The Hook
##    2:         0        25   1            1         0              The Hook
##    3:         0        37   1            0         0               The Six
##    4:         0        56   1            0         0                 The G
##    5:         0        35   1            0         0         Andy's Puzzle
##   ---
## 6638:         0        25   1            1         0                Candle
## 6639:         0        50   1            1         0               Diamond
## 6640:         0       124   1            1         0          Laughing Man
## 6641:         0        40   1            1         6 Walking Person Puzzle
## 6642:         0        35   1            0         0     House of Tangrams
##                  Timestamp TimeUsed_log factor_gender level_gender
##    1: 2017-03-03 14:58:01     3.514556             1            M
##    2: 2017-03-03 14:56:49     3.629289             1            M
##    3: 2017-03-03 07:30:34     3.628731             1            M
##    4: 2017-03-03 07:29:20     3.633420             1            M
##    5: 2017-03-03 07:26:47     3.535787             1            M
##   ---
## 6638: 2012-10-21 23:49:49     4.169297             1            M
## 6639: 2012-10-21 21:41:28     4.837250             1            M
## 6640: 2012-10-21 21:39:07     5.588701             1            M
## 6641: 2012-10-13 12:50:25     4.674155             1            M
## 6642: 2012-07-24 08:14:39     4.099962             1            M
##       factor_STEM level_stem SampleSize log_timeused_mean timeused_mean  n
##    1:           0         -1          2          3.571922      35.64350  2
##    2:           0         -1          2          3.571922      35.64350  2
##    3:           0         -1         10          4.001152      60.55480 10
##    4:           0         -1         10          4.001152      60.55480 10
##    5:           0         -1         10          4.001152      60.55480 10
##   ---
## 6638:           0         -1          5          4.633952     122.94840  5
## 6639:           0         -1          5          4.633952     122.94840  5
## 6640:           0         -1          5          4.633952     122.94840  5
## 6641:           0         -1          5          4.633952     122.94840  5
## 6642:           0         -1         70          4.197110      80.63254 50
##              sd
```

```
##    1: 0.08112837
##    2: 0.08112837
##    3: 0.45101747
##    4: 0.45101747
##    5: 0.45101747
##   ---
## 6638: 0.65355941
## 6639: 0.65355941
## 6640: 0.65355941
## 6641: 0.65355941
## 6642: 0.61283044
```

```r
gender_samplingdis <- gender_samplingdis[,c("GroupName", "level_gender", "SampleSize", "timeused_mean",
gender_samplingdis <- gender_samplingdis[!duplicated(gender_samplingdis),]

### Aggregate data for vertical lines
gender_samplingdis_vline <- gender_samplingdis %>%
  group_by(level_gender) %>%
  summarise(log_timeused_mean = mean(log_timeused_mean), timeused_mean = mean(timeused_mean))

ggplot(gender_samplingdis, aes(timeused_mean, fill = level_gender)) +
  geom_density(alpha = 0.3) +
  geom_vline(data=gender_samplingdis_vline, aes(xintercept=timeused_mean,  colour=level_gender),
             linetype="dashed", size=1) +
  labs(title = "Distribution of Sample Means of Different Gender Levels") +
  labs(x = "Sample Mean", y = "Density") +
  ylim(0.00, 0.025) +
  scale_x_continuous(minor_breaks = seq(0, 200, by = 25))
```




```
ggplot(gender_samplingdis, aes(log_timeused_mean, fill = level_gender)) +
  geom_density(alpha = 0.3) +
  geom_vline(data=gender_samplingdis_vline, aes(xintercept=log_timeused_mean,  colour=level_gender),
             linetype="dashed", size=1) +
  labs(title = "Distribution of Sample Means of Different Gender Levels") +
  labs(x = "Sample Mean (log)", y = "Density") +
  ylim(0.00, 1.00)
```


Distribution of Sample Means of Different Gender Levels
1.00
0.75
0.50
0.25
0.00
Density
3
4
5
6
Sample Mean (log)
level_gender
F
M
Others