Survey Analysis

(1) Load the datasets

```
ds <- read.csv("second-survey-js.csv", head=T, sep=",")
atom_ds <- read.csv("s02-atoms.csv", head=T, sep=",")
nrow(ds)
```

```
## [1] 1839
```

```
colnames(ds)
```

```
##  [1] "rid"        "sid"        "qid"        "atom"       "time"
##  [6] "correct"    "experience" "education"  "total"      "ref"
## [11] "empty"      "ssid"
```

```
nrow(atom_ds)
```

```
## [1] 48
```

```
colnames(atom_ds)
```

```
## [1] "atomId"          "atomDescription"
```

(2) Filter the datasets

```
# sqldf("select rid, count(*) from ds group by rid")
# sqldf("select ref, count(*) from ds group by ref")
# sqldf("select total, count(*) from ds group by total")


subSet <- ds[ds$ref=="reddit" & ds$total == "YES",]
summary(subSet$total)
```

```
##    Length     Class      Mode
##      1500 character character
```

```
squareLengths <- tapply(subSet$time,subSet$rid,length)
completeCases <- names(squareLengths)[squareLengths==24]
ds <- ds[is.element(el = ds$rid, set = completeCases),]
dim(ds)
```

```
## [1] 816  12
```

```
sqldf("select rid, count(*) from ds group by rid")
```

```
##    rid count(*)
## 1    1       24
## 2   13       24
## 3   16       24
## 4   18       24
## 5   19       24
```

```
## 6   24       24
## 7   25       24
## 8   28       24
## 9   35       24
## 10  47       24
## 11  48       24
## 12  51       24
## 13  55       24
## 14  57       24
## 15  59       24
## 16  60       24
## 17  61       24
## 18  63       24
## 19  65       24
## 20  79       24
## 21  81       24
## 22  87       24
## 23  89       24
## 24  91       24
## 25  95       24
## 26  97       24
## 27  98       24
## 28 103       24
## 29 104       24
## 30 105       24
## 31 112       24
## 32 113       24
## 33 115       24
## 34 117       24
```

(3) Exploratory Data Analysis

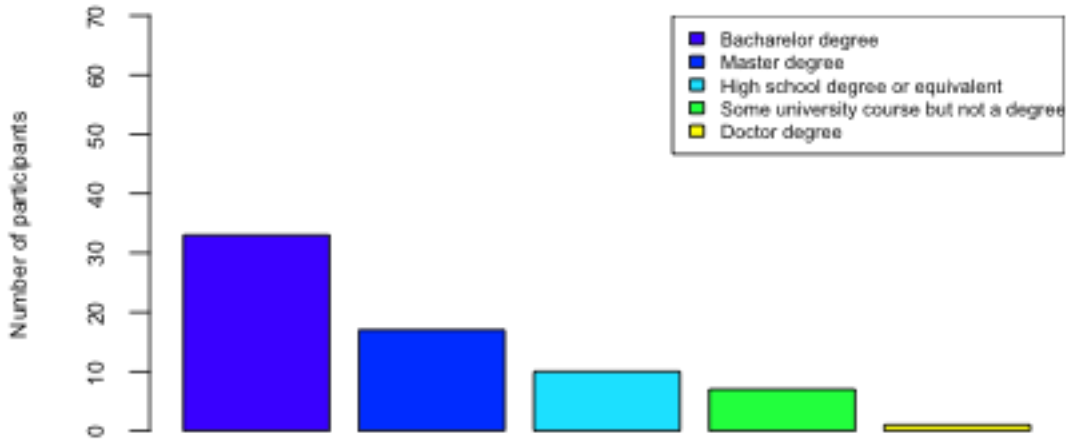
Demographics

```
educationLevelIds <- c(1, 2, 3, 4, 5)
educationLevelLabels <- c("High school degree or equivalent",
                          "Some university course but not a degree",
                          "Bacharelor degree", "Master degree",
                          "Doctor degree")

education_ds <- data.frame(educationLevelIds, educationLevelLabels)
colnames(education_ds) <- c("id", "description")

tmp <- sqldf("select distinct sid, education from ds")

demEducation <- sqldf("select b.description, count(*) total
                      from tmp a, education_ds b
                      where a.education = b.id
                      group by b.description
                      order by 2 desc")

demEducation
```

Education

```
##                              description total
## 1                       Bacharelor degree    33
## 2                           Master degree    17
## 3       High school degree or equivalent    10
## 4 Some university course but not a degree     7
## 5                           Doctor degree     1
```

```
barplot(demEducation$total, col=topo.colors(5),
        ylim = c(0, 70), cex=0.7, cex.lab = 0.7, cex.axis=0.7,
        ylab="Number of participants")

legend("topright", legend=demEducation$description, fill=topo.colors(5), cex=0.6)
```



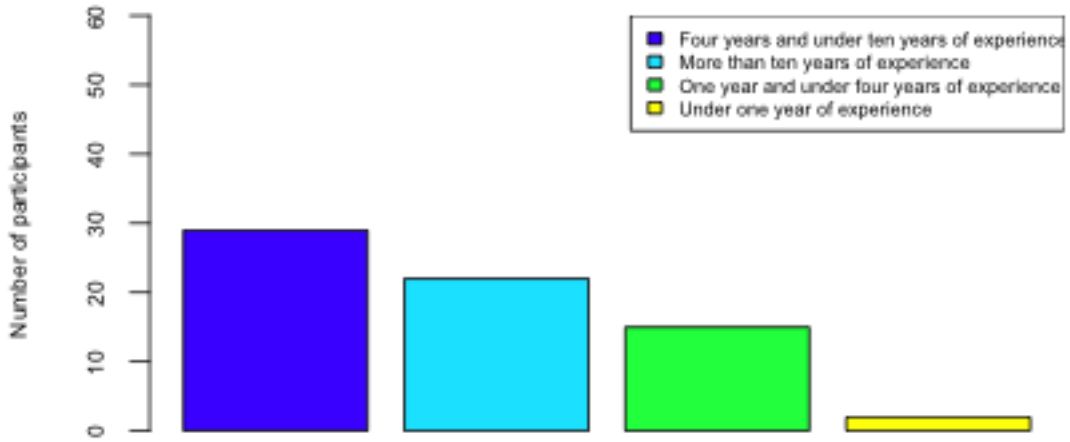

```
experienceLevelIds <- c(1, 2, 3, 4)
experienceLevelLabels <- c("Under one year of experience",
                           "One year and under four years of experience",
                           "Four years and under ten years of experience",
                           "More than ten years of experience")

experience_ds <- data.frame(experienceLevelIds, experienceLevelLabels)

colnames(experience_ds) <- c("id", "description")

tmp <- sqldf("select distinct sid, experience from ds")
demExperience <- sqldf("select experience, count(*) total
                       from tmp group by experience order by 1")

demExperience <- sqldf("select a.experience, b.description, a.total
```

```
                        from demExperience a, experience_ds b
                        where a.experience = b.id
                        order by 3 desc")
demExperience
```

Experience

```
##   experience                                    description total
## 1          3 Four years and under ten years of experience    29
## 2          4            More than ten years of experience    22
## 3          2  One year and under four years of experience    15
## 4          1                  Under one year of experience     2
```

```
barplot(demExperience$total, col=topo.colors(4),
        ylim = c(0, 60), cex=0.7, cex.lab = 0.7, cex.axis=0.7,
        ylab="Number of participants")

legend("topright", legend=demExperience$description, fill=topo.colors(4), cex=0.6)
```



Total number of correct answers (Table III)

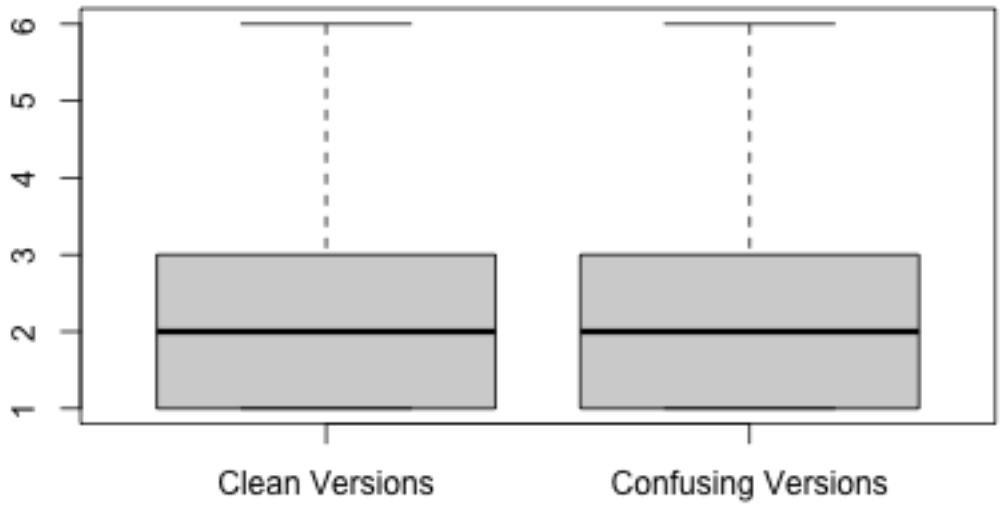
```
codeWithAtoms <- sqldf("select qid, count(*) confuseCode, avg(time) timeConfuseCode
                        from ds
                        where atom == 'YES' and correct = 'CORRECT'
                        group by qid")

codeWithoutAtoms <- sqldf("select qid, count(*) cleanCode, avg(time) timeCleanCode
                        from ds
                        where atom == 'NO' and correct = 'CORRECT'
                        group by qid")
```

```
codeWithAtoms["atomId"] = codeWithAtoms$qid %% 24
codeWithoutAtoms["atomId"] = codeWithoutAtoms$qid %% 24

merged <- sqldf("select c.atomDescription as Atom,
                        a.confuseCode as 'Confusing Versions',
                        b.cleanCode as 'Clean Versions',
                        (b.cleanCode * 100 / a.confuseCode) -100 as 'Delta (%)'
                from codeWithAtoms a, codeWithoutAtoms b, atom_ds c
                where a.atomId = b.atomId and a.atomId = c.atomId
                order by 4 desc")

xtable(merged)
```

```
## % latex table generated in R 4.2.0 by xtable 1.8-4 package
## % Wed Oct  5 10:35:23 2022
## \begin{table}[ht]
## \centering
## \begin{tabular}{rlrrr}
##   \hline
##  & Atom & Confusing Versions & Clean Versions & Delta (\%) \\
##   \hline
## 1 & Type Conversion &   2 &   6 & 200 \\
##   2 & Change Literal Encoding &   3 &   7 & 133 \\
##   3 & Comma Operator &   7 &  12 &  71 \\
##   4 & Arithmetic As Logic &   3 &   5 &  66 \\
##   5 & Indentation No Braces &   4 &   6 &  50 \\
##   6 & Assignment As Value &   5 &   7 &  40 \\
##   7 & Repurposed Variables &   3 &   4 &  33 \\
##   8 & Post Increment &  13 &  15 &  15 \\
##   9 & Arrow Function &   7 &   8 &  14 \\
##   10 & Ommited Curly Braces &  18 &  20 &  11 \\
##   11 & Array Destructuring &  18 &  20 &  11 \\
##   12 & Logic As Control Flow &  21 &  23 &   9 \\
##   13 & Indentation With Braces &  19 &  20 &   5 \\
##   14 & Infix Operator Precedence &   6 &   6 &   0 \\
##   15 & Ternary Operator &  23 &  23 &   0 \\
##   16 & Constant Variables &  22 &  21 &  -5 \\
##   17 & Dead Unreachable Repeated &  23 &  22 &  -5 \\
##   18 & Property Access &  22 &  21 &  -5 \\
##   19 & Array Spread &  25 &  21 & -16 \\
##   20 & Implicit Predicate &  21 &  15 & -29 \\
##   21 & Automatic Semicolon Insertion &   2 &   1 & -50 \\
##   22 & Object Spread &   5 &   2 & -60 \\
##    \hline
## \end{tabular}
## \end{table}
```

```
wrongAnswers = ds[ds$correct == 'WRONG', ]

wrongAnswersByStudentTreatment <- aggregate(rid~sid+atom,
                                            data = wrongAnswers,
                                            FUN=length)

boxplot(wrongAnswersByStudentTreatment$rid ~ wrongAnswersByStudentTreatment$atom
```

```
        , ylab = "", xlab = "", main = "",
        names = c("Clean Versions", "Confusing Versions"))
```



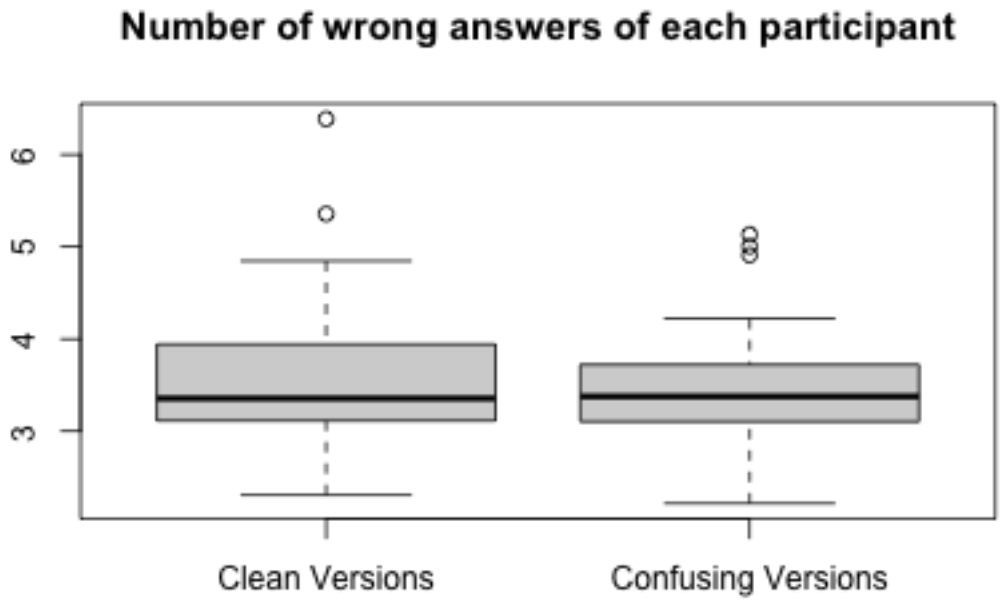
Average time for correct answers (Table IV)

```
merged <- sqldf("select c.atomDescription as Atom,
                        a.timeConfuseCode as 'Confuse Code',
                        b.timeCleanCode as 'Clean Code',
                        (b.timeCleanCode * 100 / a.timeConfuseCode) -100 as 'Delta (%)'
                from codeWithAtoms a, codeWithoutAtoms b, atom_ds c
                where a.atomId = b.atomId and a.atomId = c.atomId
                order by 4")
xtable(merged)
```

```
## % latex table generated in R 4.2.0 by xtable 1.8-4 package
## % Wed Oct  5 10:35:24 2022
## \begin{table}[ht]
## \centering
## \begin{tabular}{rlrrr}
##   \hline
##  & Atom & Confuse Code & Clean Code & Delta (\%) \\
##   \hline
## 1 & Change Literal Encoding & 82.30 & 13.36 & -83.77 \\
##   2 & Infix Operator Precedence & 43.98 & 24.24 & -44.89 \\
##   3 & Comma Operator & 64.03 & 35.84 & -44.03 \\
##   4 & Constant Variables & 24.59 & 13.98 & -43.14 \\
##   5 & Type Conversion & 50.20 & 29.76 & -40.72 \\
##   6 & Object Spread & 66.40 & 43.85 & -33.96 \\
##   7 & Indentation With Braces & 36.39 & 28.74 & -21.02 \\
```

```
##    8 & Automatic Semicolon Insertion & 63.78 & 55.77 & -12.57 \\
##    9 & Dead Unreachable Repeated & 18.50 & 17.08 & -7.66 \\
##    10 & Arrow Function & 43.55 & 40.44 & -7.15 \\
##    11 & Repurposed Variables & 69.92 & 66.76 & -4.52 \\
##    12 & Ommited Curly Braces & 25.77 & 26.07 & 1.15 \\
##    13 & Assignment As Value & 41.67 & 45.88 & 10.11 \\
##    14 & Arithmetic As Logic & 28.57 & 33.36 & 16.76 \\
##    15 & Array Destructuring & 22.48 & 28.68 & 27.59 \\
##    16 & Post Increment & 40.90 & 54.84 & 34.09 \\
##    17 & Array Spread & 39.02 & 52.47 & 34.49 \\
##    18 & Logic As Control Flow & 37.26 & 50.84 & 36.43 \\
##    19 & Indentation No Braces & 23.01 & 40.40 & 75.61 \\
##    20 & Property Access & 34.64 & 66.65 & 92.40 \\
##    21 & Ternary Operator & 27.39 & 87.85 & 220.69 \\
##    22 & Implicit Predicate & 30.12 & 252.82 & 739.29 \\
##     \hline
## \end{tabular}
## \end{table}
```

```r
correctAnswers = ds[ds$correct == 'CORRECT', ]

correctAnswersByStudentTreatment <- aggregate(time~sid+atom,
                                              data = correctAnswers,
                                              FUN=mean)

boxplot(log(correctAnswersByStudentTreatment$time) ~ correctAnswersByStudentTreatment$atom
        , ylab = "", xlab = "", main = "Number of wrong answers of each participant",
        names = c("Clean Versions", "Confusing Versions"))
```



```
prop.table(table(ds$correct,ds$atom),margin = 2)
```

```
##
##                   NO       YES
##   CORRECT 0.7181373 0.6740196
##   WRONG   0.2818627 0.3259804
```

```
chisq.test(ds$correct,ds$atom)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  ds$correct and ds$atom
## X-squared = 1.6741, df = 1, p-value = 0.1957
```

(4) Regression Analysis (correctness~atom + experience + education)

```
experience <- as.factor(ds$experience)
education <- as.factor(ds$education)
contrr<-matrix(c(rep(1,4),c(1/2,1/2,-1/2,-1/2),c(1,-1,0,0),c(0,0,1,-1)),byrow=TRUE,nrow=4)
contrasts(experience)<-solve(contrr)[,2:4]

ds$atom <- as.factor(ds$atom)
ds$correct <- as.factor(ds$correct)

mod <- glm(ds$correct~ds$atom+experience+education,family = "binomial")
summary(mod)
```

```
##
## Call:
## glm(formula = ds$correct ~ ds$atom + experience + education,
##     family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3941  -0.8035  -0.7149   1.1472   1.9437
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.01341    0.21274  -0.063   0.9497
## ds$atomYES   0.22303    0.15768   1.414   0.1572
## experience1  0.04762    0.26637   0.179   0.8581
## experience2 -0.52572    0.48657  -1.080   0.2799
## experience3 -0.22857    0.18971  -1.205   0.2283
## education2  -0.44600    0.30118  -1.481   0.1386
## education3  -1.23823    0.22200  -5.578 2.44e-08 ***
## education4  -1.30554    0.25864  -5.048 4.47e-07 ***
## education5  -1.80216    0.80437  -2.240   0.0251 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1002.29  on 815  degrees of freedom
```

```
## Residual deviance:  950.84  on 807  degrees of freedom
## AIC: 968.84
##
## Number of Fisher Scoring iterations: 4
```

```
car::Anova(mod, type=3)
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: ds$correct
##             LR Chisq Df Pr(>Chisq)
## ds$atom       2.006  1     0.1567
## experience    4.657  3     0.1987
## education    42.510  4  1.308e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Impact considering Developer Experience

```
for(i in 1:4){
  print(paste("Experience ", i ))
  print(prop.table(table(ds$correct[ds$experience==i],ds$atom[ds$experience==i]),margin=2))
  print(chisq.test(table(ds$correct[ds$experience==i],ds$atom[ds$experience==i])))
}
```

Impact considering Developer Education

```
for(i in 1:5){
  print(paste("Education ", i ))
  print(prop.table(table(ds$correct[ds$education==i],ds$atom[ds$education==i]),margin=2))
  print(chisq.test(table(ds$correct[ds$education==i],ds$atom[ds$education==i])))
}
```

Impact of Individual Atoms and Hypotheses Testing

```
chiTest <- c()
oddsRatio <- c()
oddsRatioTest <- c()
ci25 <- c()
ci975 <- c()
mannWhitneyTest <- c()
cliffDelta <- c()
for(i in 1:10){
  subSet <- ds[is.element(el = ds$qid, set = c(i,i+10)),]
  print(atomDescription[i])

  tableCorrectness <- table(subSet$correct,subSet$atom)
  tableTime <- aggregate(ds$time, by=list(atom=ds$atom), FUN=sum)

  print(tableCorrectness)
  print(tableTime)

  test <- chisq.test(tableCorrectness)
  print(test)
```

```
  chiTest[i] <- format.pval(test$p.value)
  oddsRatio[i] <- odds.ratio(tableCorrectness)$OR
  oddsRatioTest[i] <- format.pval(odds.ratio(tableCorrectness)$p)
  ci25[i] <- odds.ratio(tableCorrectness, level=0.95)$"2.5 %"
  ci975[i] <- odds.ratio(tableCorrectness, level=0.95)$"97.5 %"

  mannWhitneyTest[i] <- format.pval(wilcox.test(subSet$time~as.factor(subSet$atom))$p.value)
  cliffDelta[i] <- cliff.delta(subSet$time~as.factor(subSet$atom))$estimate

  experience <- as.factor(subSet$experience)
  contrr<-matrix(c(rep(1,4),c(1/2,1/2,-1/2,-1/2),c(1,-1,0,0),c(0,0,1,-1)),byrow=TRUE,nrow=4)
  contrasts(experience)<-solve(contrr)[,2:4]

  mod <- glm(subSet$correct ~ subSet$atom +  experience,family="binomial")
  print(summary(mod))
  print(prop.table(table(subSet$correct,subSet$experience),margin = 2))
}
analysis_df <- data.frame(atomDescription, chiTest, oddsRatio,
                          oddsRatioTest, ci25, ci975, mannWhitneyTest, cliffDelta)

analysis_df <- analysis_df[order(-oddsRatio),]

colnames(analysis_df) <- c("Atom", "ChiTest",
                           "Odds Ratio Correctness",
                           "p-value", "CI 2.5%", "CI 97.5%",
                           "Wilcox Test (Time)",
                           "Cliff Delta")

xtable(analysis_df)
```