tar2

Part 1:

```
data <- read.xlsx("data_and_headers_processed.xlsx", 1, stringsAsFactors=T)
```

2. Possible problems:

- Hebrew mixed with English - we took this problem and fixed the input file to include only english letters
- The sex, age feature have missing values - removing this records

```
    data$Age <- as.numeric(as.character(data$Age))
    data <- data[!is.na(data$Age),]
    data <- data[!is.na(data$Sex),]
    summary(data$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   25.25   28.00   31.29   33.00   75.00
```

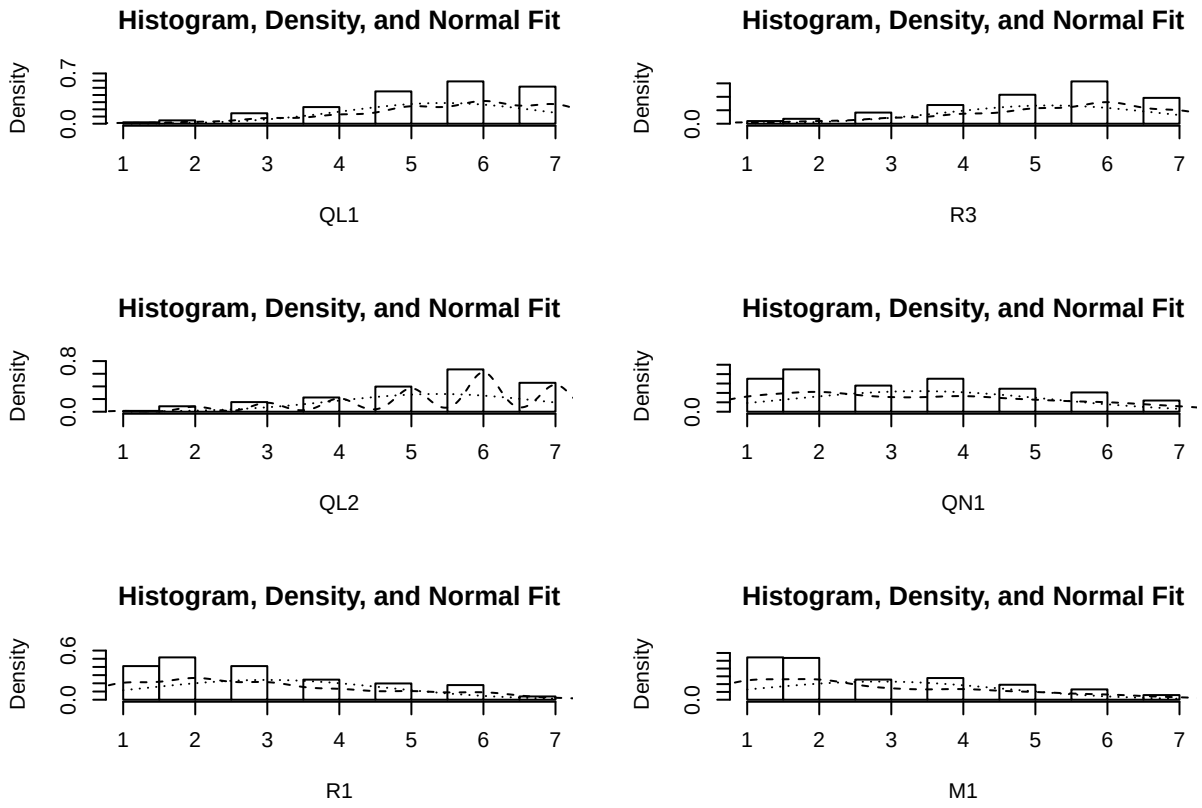
```
    summary(data$Sex)
```

```
##  C1  C2
## 183 119
```

- The features don't really distribute noramly (i.e according to normal distribution) - it's not a problem by unless we assume it should distribute normaly. Here are few exmaples

```
    numeric.feature.names <- names(data)[c(which(names(data)=='QL1'):which(names(data)=='DI1'), which(na

    multi.hist(x = data[,numeric.feature.names[1:6]])
```



Clean data set - remove NA Age values `{ r warning=FALSE, message=FALSE}`

Create clarity, politeness, satisfaction variables

```
data.for.clarity <- cbind(data[,c("C1", "C2", "C3", "C5")], 8-data$C4, 8-data$C6)
clarity <- apply(data.for.clarity, MARGIN = 1, FUN = mean)

data.for.politeness <- cbind(data[,c("P1", "P2", "P4", "P5", "P6")], 8-data$P3)
politeness <- apply(data.for.politeness, MARGIN = 1, FUN = mean)

data.for.satisfaction <- cbind(data[,c("S1", "S2", "S3", "S5", "S6")], 8-data$S4)
satisfaction <- apply(data.for.satisfaction, MARGIN = 1, FUN = mean)

#now adding them to the data frame
data <- cbind(data, clarity = clarity, politeness = politeness, satisfaction = satisfaction)
```

Part 2

Part 2.1

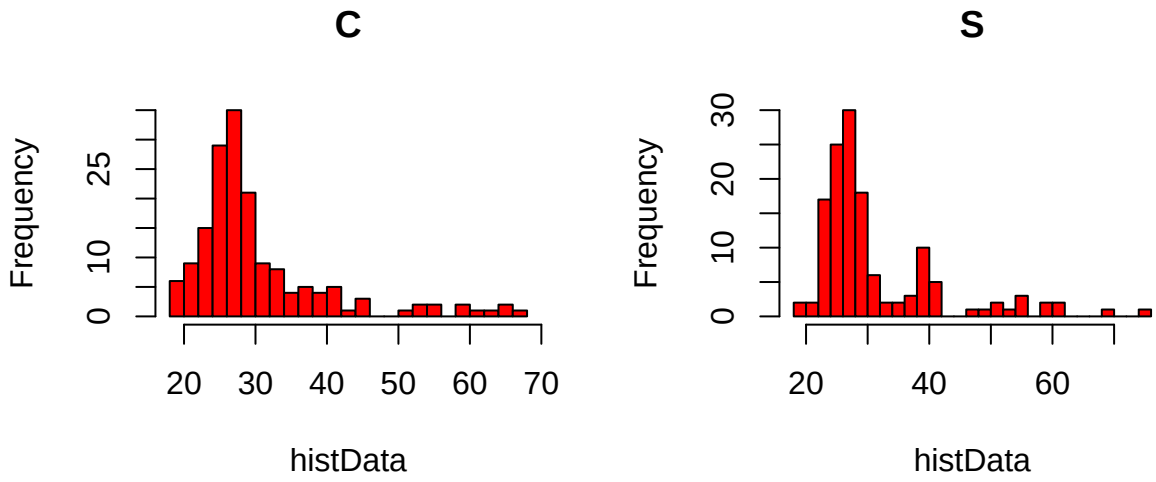
Descriptive statistics Comparing the different parameters grouped by system type.

data$Age

```
par(mfrow=c(1,2))
combineSummaryFrame(data[data$System=='C',]$Age, data[data$System=='S',]$Age, rowNames = c('C', 'S'))
```

```
##   Min. 1st Qu. Median  Mean 3rd Qu. Max.
## C   18      25     28 30.85   32.75   67
## S   19      26     28 31.82   34.25   75
```

```
invisible(drawHist(data[data$System=='C',]$Age, br=20, main='C')) #suppress ## NULL
invisible(drawHist(data[data$System=='S',]$Age, br=20, main='S'))
```



data$Sex

```
par(mfrow=c(1,2))
colnames <- c("men", "women")
combineSummaryFrame(data[data$System=='C',]$Sex, data[data$System=='S',]$Sex, colnames = colnames, rowNa
```

```
##   men women
## C 100    66
## S  83    53
```

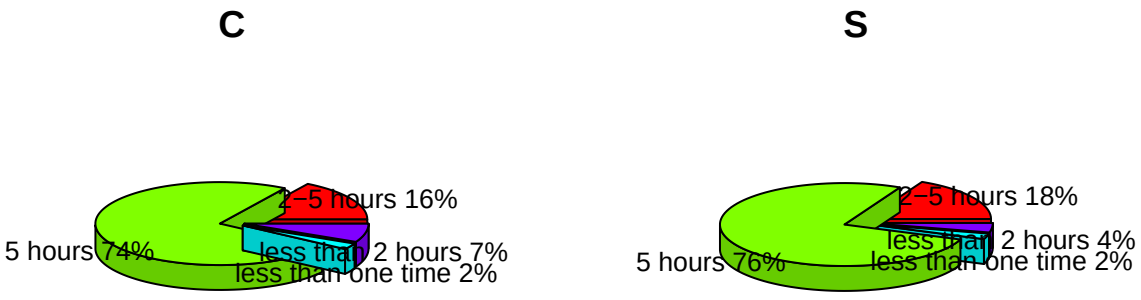
```
invisible(drawPieChart(table(data[data$System=='C',]$Sex), colnames, main='C'))
invisible(drawPieChart(table(data[data$System=='S',]$Sex), colnames, main='S'))
```

data$Comp_Use_Freq

```
par(mfrow=c(1,2))
colnames <- c("2-5 hours", "5 hours", "less than one time", "less than 2 hours")
combineSummaryFrame(data[data$System=='C',]$Comp_Use_Freq, data[data$System=='S',]$Comp_Use_Freq, colnar
```

```
##   2-5 hours 5 hours less than one time less than 2 hours
## C        27     123                  4                12
## S        24     104                  3                 5
```

```
invisible(drawPieChart(table(data[data$System=='C',]$Comp_Use_Freq), colnames, main='C'))
invisible(drawPieChart(table(data[data$System=='S',]$Comp_Use_Freq), colnames, main='S'))
```

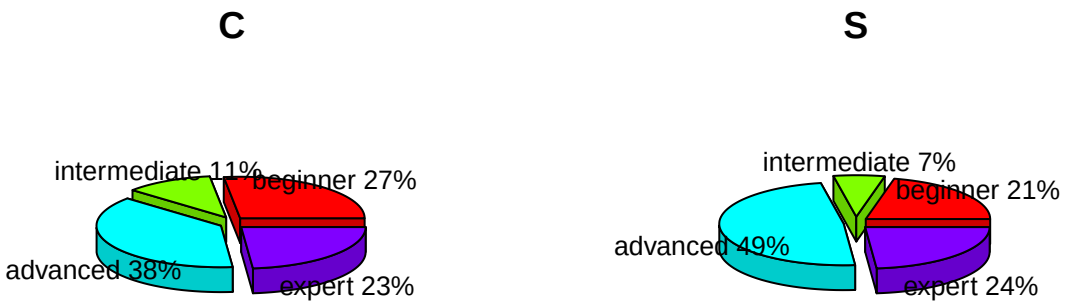


data$Comp_Use_Know

```
par(mfrow=c(1,2))
colnames <- c("beginner", "intermediate", "advanced",  "expert")
combineSummaryFrame(data[data$System=='C',]$Comp_Use_Know, data[data$System=='S',]$Comp_Use_Know, colnar
```

```
##   beginner intermediate advanced expert
## C       45           19       63     39
## S       29            9       66     32
```

```
invisible(drawPieChart(table(data[data$System=='C',]$Comp_Use_Know), colnames, main='C'))
invisible(drawPieChart(table(data[data$System=='S',]$Comp_Use_Know), colnames, main='S'))
```

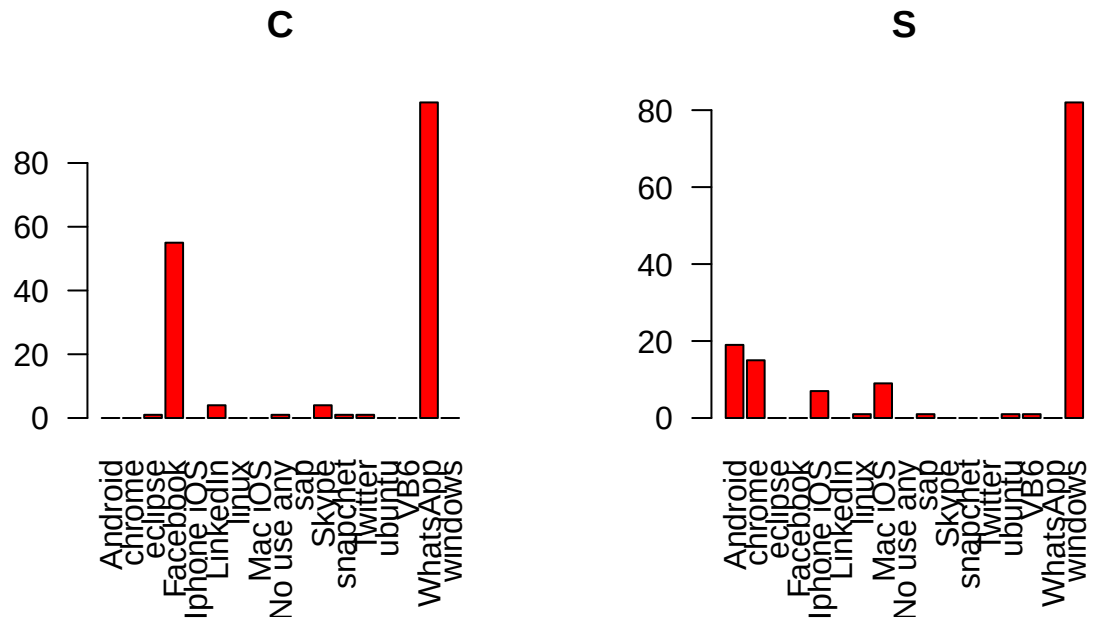


data$Selected_Software

```
par(mfrow=c(1,2))
combineSummaryFrame(data[data$System=='C',]$Selected_Software, data[data$System=='S',]$Selected_Softwar
```

```
##   Android chrome eclipse Facebook Iphone iOS LinkedIn linux Mac iOS
## C       0      0       1       55          0        4     0       0
## S      19     15       0        0          7        0     1       9
##   No use any sap Skype snapchet Twitter ubuntu VB6 WhatsApp windows
## C          1   0     4        1       1      0   0       99       0
## S          0   1     0        0       0      1   1        0      82
```

```
invisible(barplot(table(data[data$System=='C',]$Selected_Software), las=2, col = 'red', main='C'))
invisible(barplot(table(data[data$System=='S',]$Selected_Software), las=2, col = 'red', main='S'))
```



Part 2.2

Create descriptive statistics for clarity, politeness, satisfactions clarity is right skewed, politeness and satisfactions are almost not skewed All variables have strong positive kurtosis

```
data_filtered <- data[data$System == 'C' & data$Age  >= 18 & data$Age<=49,]

stat_data <- data_filtered[ ,names(data_filtered) %in% c("clarity", "politeness", "satisfaction")]
stat_res <- data.frame(
  apply(stat_data, 2, length),
  apply(stat_data, 2, mean, na.rm=TRUE),
  apply(stat_data, 2, sd, na.rm=TRUE),
  apply(stat_data, 2, min, na.rm=TRUE),
  apply(stat_data, 2, max, na.rm=TRUE),
  apply(stat_data, 2, kurtosis, na.rm=TRUE),
  apply(stat_data, 2, skewness, na.rm=TRUE)
)
colnames(stat_res) <- c('count', 'mean', 'sd', 'min', 'max', 'kurtosis', 'skewness')
stat_res
```

```
##              count     mean        sd      min max kurtosis    skewness
## clarity        154 5.408009 0.9030816 3.000000   7 2.589470 -0.29878680
## politeness     154 4.656926 1.0948544 1.666667   7 2.747956 -0.06905606
## satisfaction   154 5.146104 0.9488226 3.000000   7 2.401156 -0.07127015
```

Part 2.3

Create correlation table for Age, clarity, politeness, satisfactions From the results can be seen that clarity, politeness and satisfactions are mutually strongly positively correlated There is some degree of negative correlation between age and satisfaction

```
stat_data_2.3 <- data_filtered[ ,names(data_filtered) %in% c("Age", "clarity", "politeness", "satisfacti
corstarsl(stat_data_2.3)
```

```
##                  Age  clarity politeness
## Age
## clarity       -0.15
## politeness    -0.05   0.51***
## satisfaction -0.19*   0.74***    0.64***
```

Part 2.4

Genarate linear regression for satisfaction depended by Age and Sex - (lmodel1), Age, Sex, clarity, politeness (lmodel2) Model is significant in both cases , but 'Coefficient of determination' (R^2 and adjusted R^2) are far greater in second case Also 'Age' and 'Sex' have no significant influence on regression model in second case

```
lmodel1 = lm(satisfaction ~ Age+Sex, data = data_filtered)
summary(lmodel1)
```

```
##
## Call:
## lm(formula = satisfaction ~ Age + Sex, data = data_filtered)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.44683 -0.61155  0.01074  0.70608  2.07916
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.93795    0.38327  15.493   <2e-16 ***
## Age         -0.03178    0.01297  -2.450   0.0154 *
## SexC2        0.30350    0.15288   1.985   0.0489 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9252 on 151 degrees of freedom
## Multiple R-squared:  0.06159,    Adjusted R-squared:  0.04916
## F-statistic: 4.956 on 2 and 151 DF,  p-value: 0.008232
```

```
lmodel2 = lm(satisfaction ~ Age+Sex+clarity+politeness, data = data_filtered)
summary(lmodel2)
```

```
##
## Call:
## lm(formula = satisfaction ~ Age + Sex + clarity + politeness,
##     data = data_filtered)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.39337 -0.28880 -0.00726  0.40600  1.66216
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.10289    0.39514   2.791  0.00594 **
## Age         -0.01590    0.00810  -1.964  0.05144 .
## SexC2        0.13132    0.09544   1.376  0.17088
## clarity      0.55307    0.06081   9.095 5.58e-16 ***
## politeness   0.31275    0.04922   6.354 2.41e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5712 on 149 degrees of freedom
## Multiple R-squared:  0.6471, Adjusted R-squared:  0.6376
## F-statistic:  68.3 on 4 and 149 DF,  p-value: < 2.2e-16
```

Generating standardized regression coefficients The standardized values correspond with previous results: * Age and Satisfaction are loosely negative correlated * 'Clarity' and 'Politeness' have far stronger influence on 'Satifaction' than 'Age' and 'Sex' * At second regression 'Age' and 'Sex' can be removed from regression formula

```
lm.beta(lmodel1)
```

```
##        Age      SexC2
## -0.1931692  0.1564973
```

```
lm.beta(lmodel2)
```

```
##         Age       SexC2     clarity  politeness
## -0.09666319  0.06771529  0.52640468  0.36088950
```

Comapare 2 models using anova test The results show that second model is significally better than the first

```
anova(lmodel1, lmodel2)
```

```
## Analysis of Variance Table
##
## Model 1: satisfaction ~ Age + Sex
## Model 2: satisfaction ~ Age + Sex + clarity + politeness
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    151 129.256
## 2    149  48.611  2    80.645 123.59 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```