STEM students perception of bioinformatics

```
library(RColorBrewer)
library(wordcloud2)
library(ggplot2)
```

Import data

Read in data and split dataframe between STEM and Bioinformatics students

```
survey<-read.csv2("../../data/BioinformaticsSurvey2022.csv",sep=',')
survey_bioinf<-survey[survey$Is.bioinformatics.the.main.focus.of.your.studies.=='Yes',c(1:7,26:77)]
survey_STEM<-survey[survey$Is.bioinformatics.the.main.focus.of.your.studies.=='No',c(1:25,74:77)]
head(survey_STEM)
```

```
##                    Timestamp How.old.are.you. What.are.your.pronouns.
## 2  2022/03/24 4:44:54 pm CET           23-26                  He/him
## 6  2022/03/24 5:34:53 pm CET           27-30                  He/him
## 7  2022/03/24 5:45:54 pm CET           23-26                 She/her
## 9  2022/03/24 5:50:29 pm CET           27-30                  He/him
## 10 2022/03/24 6:15:55 pm CET           23-26                  He/him
## 11 2022/03/24 6:19:35 pm CET           23-26                  He/him
##                                                                    Where.are.you.from.
## 2  Northern-East Italy (Trentino-Alto Adige, Veneto, Friuli-Venezia Giulia, Emilia-Romagna)
## 6                  Southern Italy (Abruzzo, Molise, Campania, Puglia, Basilicata, Calabria)
## 7                                          Central Italy (Toscana, Umbria, Lazio, Marche)
## 9  Northern-East Italy (Trentino-Alto Adige, Veneto, Friuli-Venezia Giulia, Emilia-Romagna)
## 10                       Northern-West Italy (Valle d'Aosta, Liguria, Lombardia, Piemonte)
## 11                                         Central Italy (Toscana, Umbria, Lazio, Marche)
##                                                                    Where.in.Italy.are.you.studying.did.you.study.
## 2  Northern-East Italy (Trentino-Alto Adige, Veneto, Friuli-Venezia Giulia, Emilia-Romagna)
## 6  Northern-East Italy (Trentino-Alto Adige, Veneto, Friuli-Venezia Giulia, Emilia-Romagna)
## 7                                          Central Italy (Toscana, Umbria, Lazio, Marche)
## 9  Northern-East Italy (Trentino-Alto Adige, Veneto, Friuli-Venezia Giulia, Emilia-Romagna)
## 10                       Northern-West Italy (Valle d'Aosta, Liguria, Lombardia, Piemonte)
## 11                                         Central Italy (Toscana, Umbria, Lazio, Marche)
##    Are.you.an.off.site.student.              What.is.your.current.position.
## 2                           Yes                           Master's student
## 6                           Yes                           Master's student
## 7                            No                           Master's student
## 9                            No PhD student (either in academia or industry)
## 10                           No PhD student (either in academia or industry)
## 11                           No                           Master's student
##    Is.bioinformatics.the.main.focus.of.your.studies.
## 2                                                 No
## 6                                                 No
## 7                                                 No
## 9                                                 No
```

```
## 10                                                 No
## 11                                                 No
##    What.is.your.current.degree.area..Please.select.the.closest.answer.that.applies..
## 2                                                                    Computer Science
## 6                                          Life Sciences (Biology/Biotechnology)
## 7                                          Life Sciences (Biology/Biotechnology)
## 9                                          Life Sciences (Biology/Biotechnology)
## 10                                         Life Sciences (Biology/Biotechnology)
## 11                                         Life Sciences (Biology/Biotechnology)
##    Have.you.ever.heard.about.bioinformatics.
## 2           Yes, I'm familiar with the term
## 6           Yes, I'm familiar with the term
## 7           Yes, I'm familiar with the term
## 9           Yes, I'm familiar with the term
## 10          Yes, I'm familiar with the term
## 11          Yes, I'm familiar with the term
##    How.would.you.describe.bioinformatics.in.3.words..Please..write.3.words.separated.by.a.comma.in.s
## 2
## 6
## 7
## 9
## 10
## 11
##    Have.you.ever.taken.a.course.in.computational.biology.bioinformatics.
## 2                                                                    Yes
## 6                                                                    Yes
## 7                                                                     No
## 9                                                                     No
## 10                                                                   Yes
## 11                                                                   Yes
##    In.your.opinion..on.a.scale.of.1..very.improbable..to.5..most.likely...where.does.a.bioinformatic
## 2
## 6
## 7
## 9
## 10
## 11
##    In.your.opinion..on.a.scale.of.1..very.improbable..to.5..most.likely...where.does.a.bioinformatic
## 2
## 6
## 7
## 9
## 10
## 11
##    In.your.opinion..on.a.scale.of.1..very.improbable..to.5..most.likely...where.does.a.bioinformatic
## 2
## 6
## 7
## 9
## 10
## 11
##    In.your.opinion..on.a.scale.of.1..very.improbable..to.5..most.likely...where.does.a.bioinformatic
## 2
## 6
```

```
## 7
## 9
## 10
## 11
##    In.your.opinion..on.a.scale.of.1..very.improbable..to.5..most.likely...where.does.a.bioinformatic
## 2
## 6
## 7
## 9
## 10
## 11
##    In.your.opinion..on.a.scale.of.1..very.improbable..to.5..most.likely...where.does.a.bioinformatic
## 2
## 6
## 7
## 9
## 10
## 11
##    In.your.opinion..on.a.scale.of.1..very.improbable..to.5..most.likely...where.does.a.bioinformatic
## 2
## 6
## 7
## 9
## 10
## 11
##    In.your.opinion..on.a.scale.of.1..very.improbable..to.5..most.likely...where.does.a.bioinformatic
## 2
## 6
## 7
## 9
## 10
## 11
##    On.a.scale.of.1..very.little..to.5..a.lot...how.much.do.you.think.bioinformatics.is.involved.in.t
## 2
## 6
## 7
## 9
## 10
## 11
##    On.a.scale.of.1..very.little..to.5..a.lot...how.much.do.you.think.bioinformatics.is.involved.in.t
## 2
## 6
## 7
## 9
## 10
## 11
##    On.a.scale.of.1..very.little..to.5..a.lot...how.much.do.you.think.bioinformatics.is.involved.in.t
## 2
## 6
## 7
## 9
## 10
## 11
##    On.a.scale.of.1..very.little..to.5..a.lot...how.much.do.you.think.bioinformatics.is.involved.in.t
```

```
## 2
## 6
## 7
## 9
## 10
## 11
##    On.a.scale.of.1..very.little..to.5..a.lot...how.much.do.you.think.bioinformatics.is.involved.in.t
## 2
## 6
## 7
## 9
## 10
## 11
##    How.did.you.come.across.this.survey.
## 2            RSG-Italy Telegram channel
## 6               From friends/colleagues
## 7               From friends/colleagues
## 9               From friends/colleagues
## 10              From friends/colleagues
## 11              From friends/colleagues
##    Had.you.heard.about.ISCB.before.this.survey.
## 2                                            No
## 6                                           Yes
## 7                                            No
## 9                                           Yes
## 10                                           No
## 11                                           No
##    Had.you.heard.about.RSG.Italy.before.this.survey.
## 2                                                 Yes
## 6                                                 Yes
## 7                                                  No
## 9                                                 Yes
## 10                                                 No
## 11                                                 No
##
## 2  In my opinion, there should be more bachelors dedicated to quantitative biology like the Genomics
## 6
## 7
## 9
## 10
## 11
```

Word cloud

Create a word-cloud with the terms used by STEM students to describe bioinformatics

```
# extract words
words<-unlist(strsplit(survey$How.would.you.describe.bioinformatics.in.3.words..Please..write.3.words.s

# preprocess
words<-trimws(words)
words<-tolower(words)
#table(words)
```

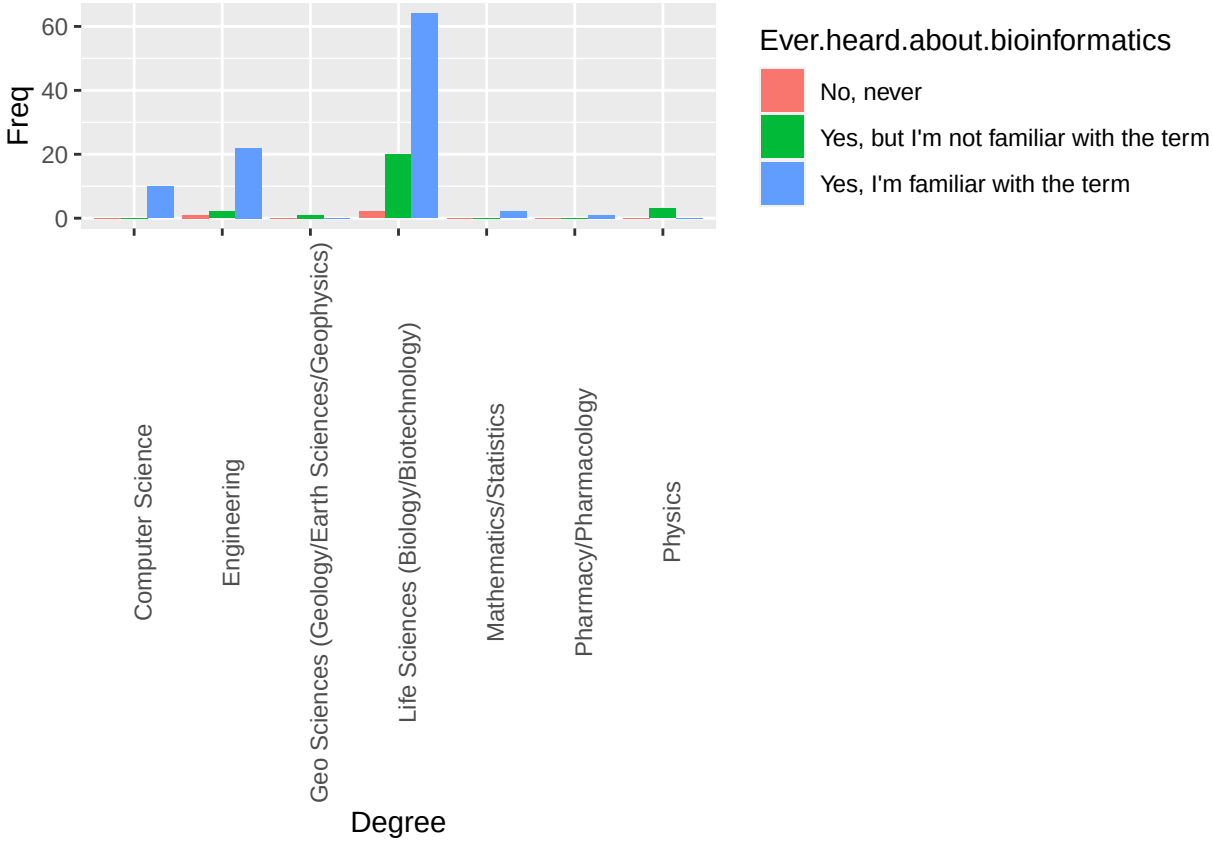
```
# correct words
tmp<-words[28]
words<-words[-28]
words<-c(words,trimws(unlist(strsplit(tmp,'/'))))
words<-words[words!="the use of a informatica tool to understand anything bio related"]
words[grep('machine',words)]<-'machine-learning'
words[grep('single',words)]<-'single-cell'
words[grep('big',words)]<-'big-data'
words[grep('rapid',words)]<-'rapid'

words<-unlist(strsplit(words,' '))
words[grep('cha',words)]<-'challenging'
words[grep('stat',words)]<-'statistics'
words[grep('useful',words)]<-'useful'
words[grep('utile',words)]<-'useful'
words[grep('model',words)]<-'modeling'
words[grep('analisis',words)]<-'analysis'
words[grep('anal',words)]<-'analysis'
words[grep('tool',words)]<-'tools'
words[grep('programs',words)]<-'tools'
words[grep('seq',words)]<-'sequencing'
words[grep('algorithm',words)]<-'algorithm'
words[grep('science',words)]<-'science'
words[grep('powerfull',words)]<-'powerful'
words[grep('confusa',words)]<-'confusing'
words[grep('programmazione',words)]<-'programming'
words[grep('moderna',words)]<-'modern'
words[grep('pred',words)]<-'prediction'
words[grep('informatic',words)]<-'informatics'
words[grep('precis',words)]<-'precise'
words[grep('sistemi',words)]<-'systems'
words[grep('pratic',words)]<-'practical'
words[grep('found',words)]<-'fundamental'
words[grep('intere',words)]<-'interesting'
words[grep('intesting',words)]<-'interesting'
words[grep('biol',words)]<-'biology'
words[grep('visulization',words)]<-'visualization'
words[grep('helpful',words)]<-'helpful'
words[grep('innov',words)]<-'innovative'
words[grep('complex',words)]<-'complexity'
words[grep('fondamental',words)]<-'fundamental'
words[grep('computation',words)]<-'computation'
words[grep('database',words)]<-'databases'
words[grep('comunication',words)]<-'communication'
words[grep('genetic',words)]<-'genetics'
words[grep('computing',words)]<-'computation'
words[grep('geomes',words)]<-'genomes'
words[grep('indispensabile',words)]<-'indispensable'
words[words%in%'omic']<-'omics'
words<-words[!words%in%c('a','in','very','the','i','and','of','to','for','know','dont',"don't")]
```

```
#table(words)
wordcloud2(data.frame(table(words)),shape='circle',color='random-dark',shuffle=F,size=1.2)
```

STEM student's knowledge on bioinformatics (stratified by degree ares)

We plot for each degree area whether the students heard about bioinformatics

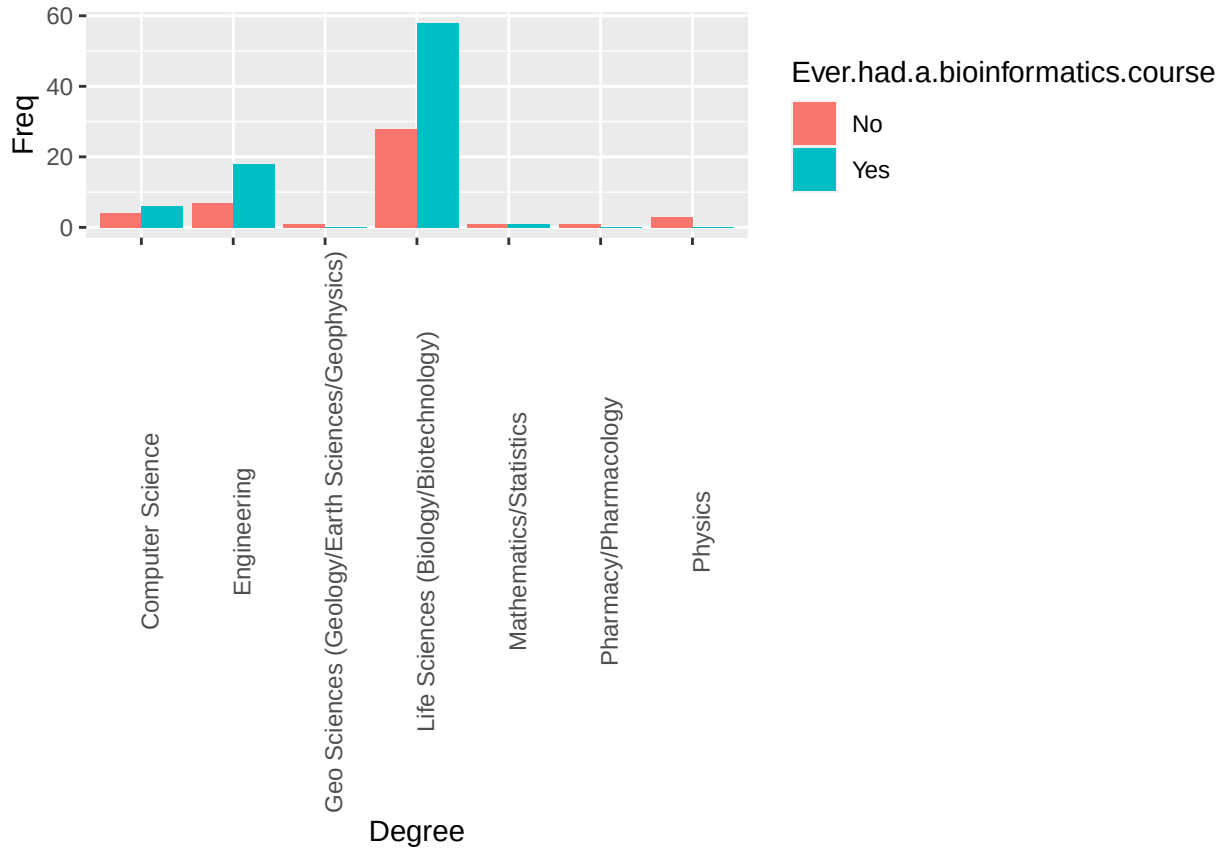
```
df<-survey_STEM[,c("What.is.your.current.degree.area..Please.select.the.closest.answer.that.applies..",
                   "Have.you.ever.heard.about.bioinformatics.")]
colnames(df)<-c('Degree','Ever heard about bioinformatics')
df<-as.data.frame(table(df))
ggplot(df,aes(fill=Ever.heard.about.bioinformatics,
              y=Freq,
              x=Degree)) +geom_bar(position='dodge',stat='identity')   +theme(axis.text.x = element_tex
```



STEM student's bioinformatics courses (stratified by degree area)

We plot for each degree area whether the students ever had a bioinformatics course

```
df<-survey_STEM[,c("What.is.your.current.degree.area..Please.select.the.closest.answer.that.applies..",
                   "Have.you.ever.taken.a.course.in.computational.biology.bioinformatics.")]
colnames(df)<-c('Degree','Ever had a bioinformatics course')
df<-as.data.frame(table(df))
ggplot(df,aes(fill=Ever.had.a.bioinformatics.course,
              y=Freq,
              x=Degree)) +geom_bar(position='dodge',stat='identity')   +theme(axis.text.x = element_tex
```



Perception of workplace between STEM and bioinformaticians

We investigate the relationship between the perception of STEM students of where a bioinformatician works and where bioinformatics students see themwelves working in the future

```
workplace<-c('Academia','Industry (R&D)', 'Data Analyst','IRCSS','Communication','Teaching','Entepreneu
STEM_indexes<-13:20
BIOINFO_indexes<-STEM_indexes+33

STEM_workplace<-survey_STEM[,STEM_indexes]
colnames(STEM_workplace)<-workplace

BIOINFO_workplace<-survey_bioinf[,BIOINFO_indexes]
colnames(BIOINFO_workplace)<-workplace
```

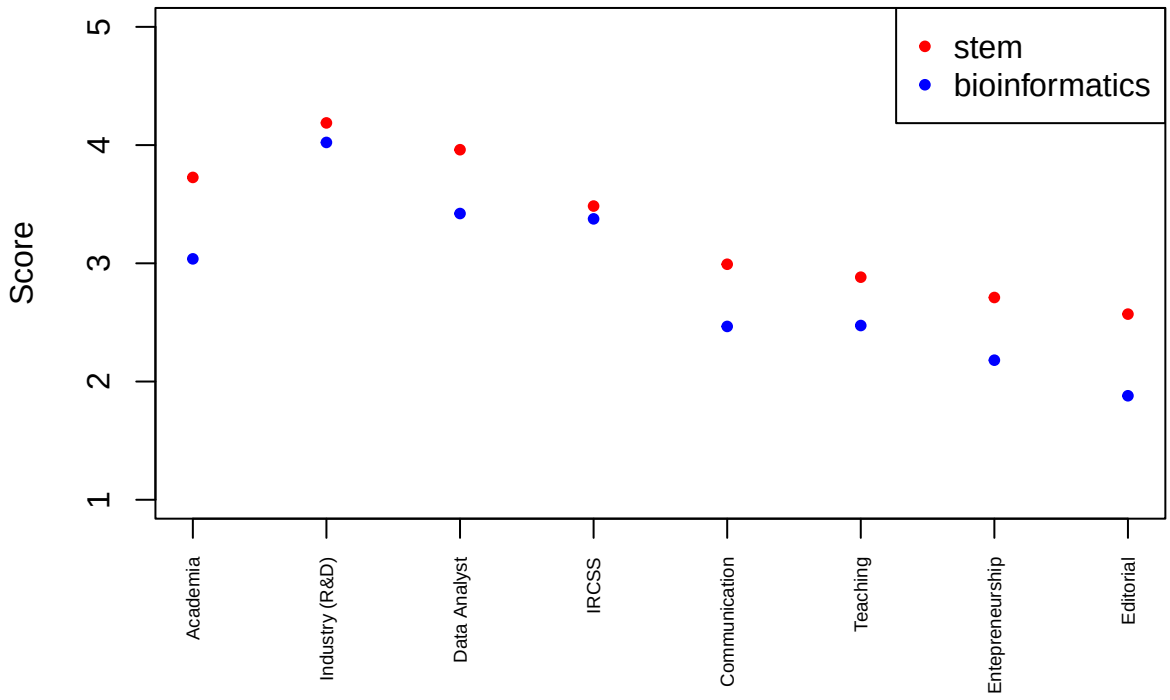
Plot the average score the two categories gave to different workplaces

```
avg_stem_workplace<-colMeans(STEM_workplace)
avg_bioinfo_workplace<-colMeans(BIOINFO_workplace)
sprintf('Correlation of average score: %f', cor(avg_stem_workplace,avg_bioinfo_workplace))
```

```
## [1] "Correlation of average score: 0.961708"
```

```
plot(x=rep(c(1:8),2),y=c(avg_stem_workplace,avg_bioinfo_workplace), xlab = '', ylab='Score', ylim=c(1,5)
axis(2)
box()
axis(1, at=1:8, labels = workplace, las=2, cex.axis=0.6)
```

```
legend('topright',legend=c('stem','bioinformatics'),col=c('red','blue'),pch=20)
```



For each workplace, we compute how many people gave scores from 1-5 in the two categories (stem/bioinformatics) and compute the correlation between STEM and bioinformatics students in each workplace (NOT SURE THIS MAKES SENSE. Should it be normalized? Does it make sense to compute correlation between 5 data points?)

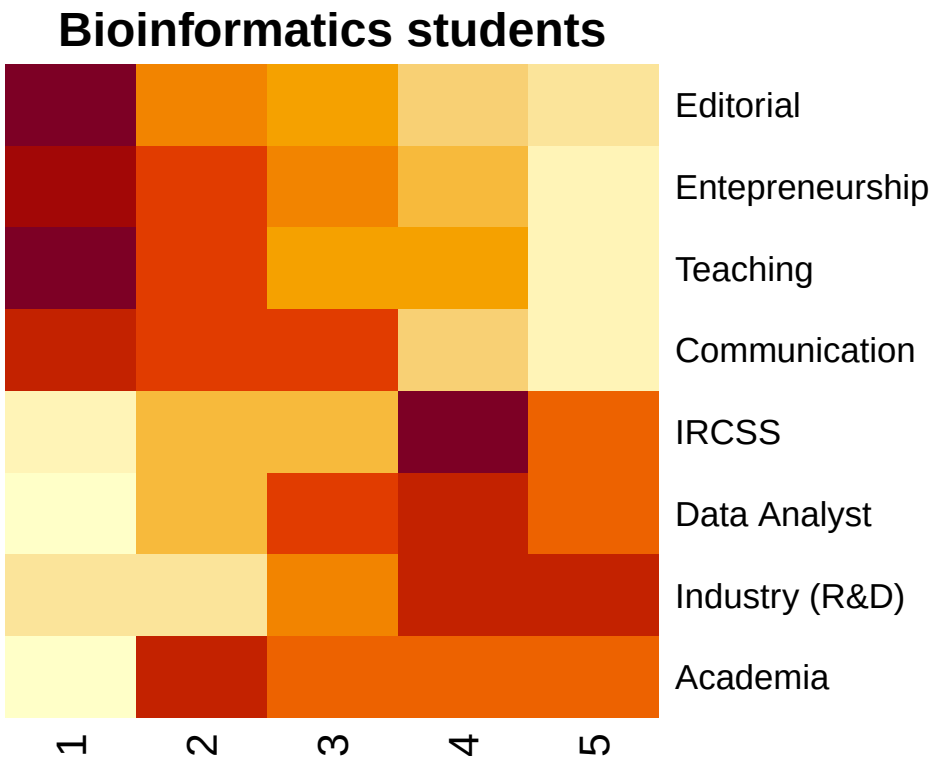
```
correlation<-c()
heatmap_stem<-c()
heatmap_bioinfo<-c()
for (i in 1:length(workplace)){
  s<-c()
  b<-c()
  for (n in 1:5){
  s[[n]]<-length(which(STEM_workplace[,i]==n))
  b[[n]]<-length(which(BIOINFO_workplace[,i]==n))
  }
  correlation[[workplace[i]]]<-cor(unlist(s),unlist(b))
  heatmap_stem<-rbind(heatmap_stem,unlist(s))
  heatmap_bioinfo<-rbind(heatmap_bioinfo,unlist(b))
}

rownames(heatmap_bioinfo)<-workplace
rownames(heatmap_stem)<-workplace

colnames(heatmap_bioinfo)<-c(1:5)
```

```
colnames(heatmap_stem)<-c(1:5)

correlation
```

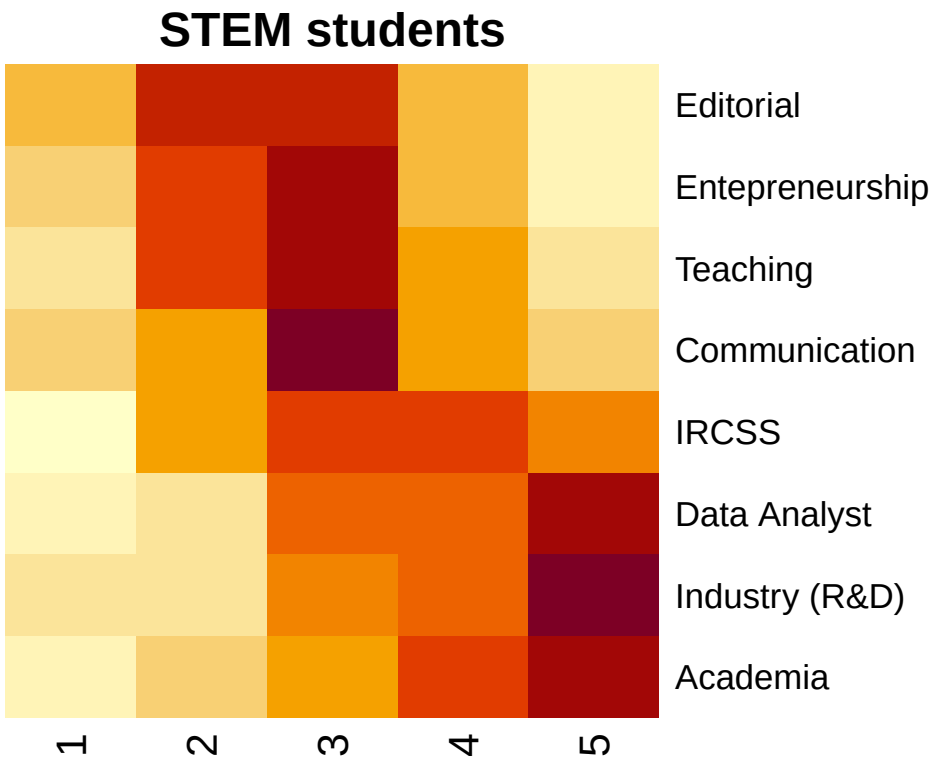
```
## $Academia
## [1] 0.4686853
##
## $`Industry (R&D)`
## [1] 0.9433833
##
## $`Data Analyst`
## [1] 0.7680763
##
## $IRCSS
## [1] 0.7537398
##
## $Communication
## [1] 0.4007069
##
## $Teaching
## [1] 0.05519182
##
## $Entepreneurship
## [1] 0.3636715
##
## $Editorial
## [1] 0.2802067
```

Plot how many people gave scores from 1-5 in the two categories

```
heatmap(heatmap_bioinfo, Rowv = NA, Colv = NA, main = 'Bioinformatics students')
```




```
heatmap(heatmap_stem, Rowv = NA, Colv = NA, main = 'STEM students')
```

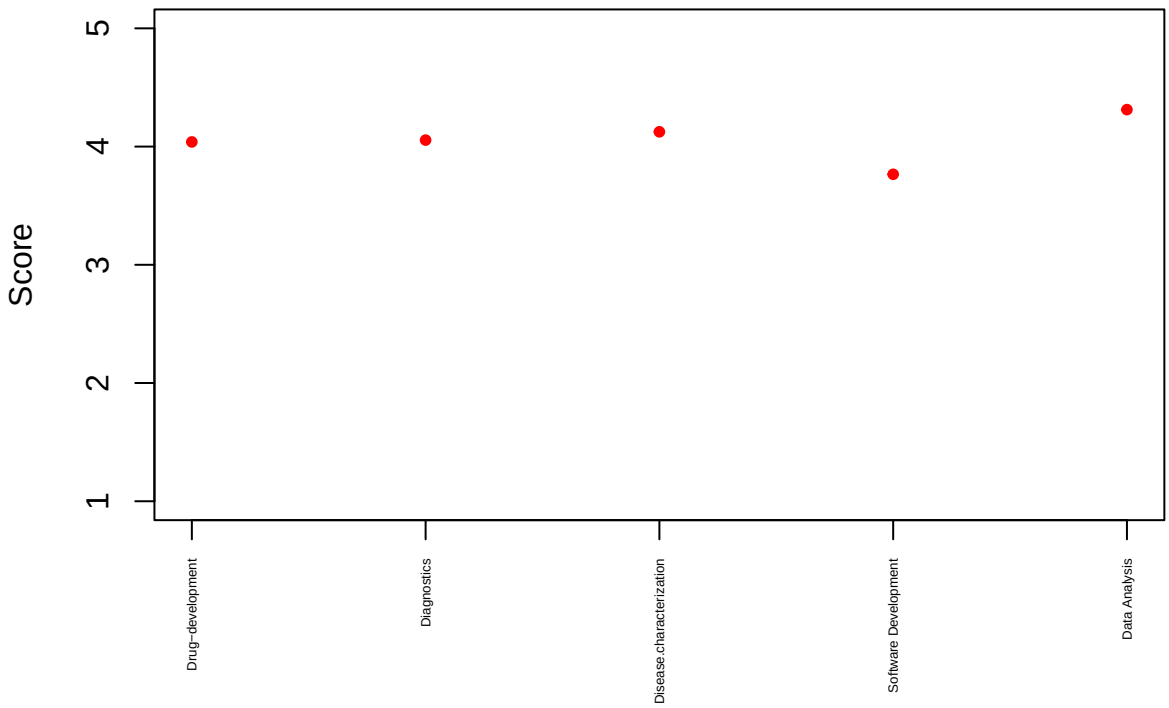


Perception of bioinformatician activities

We visualize the average score given by STEM students to different activities

```
activity<-c('Drug-development','Diagnostics', 'Disease.characterization','Software Development','Data An
activity_indexes<-21:25
STEM_activity<-survey_STEM[,activity_indexes]
colnames(STEM_activity)<-activity
```

```
# average
avg_stem_activity<-colMeans(STEM_activity)
plot(avg_stem_activity, xlab = '', ylab='Score', ylim=c(1,5),axes = FALSE, col='red',pch=20)
axis(2)
box()
axis(1, at=1:5, labels = activity, las=2, cex.axis=0.4)
```



We visualize the number of votes given to each score and activity

```
heatmap_activity<-c()
for (i in 1:length(activity)){
  s<-c()
  for (n in 1:5){
    s[[n]]<-length(which(STEM_activity[,i]==n))
  }
  heatmap_activity<-rbind(heatmap_activity,unlist(s))
}

rownames(heatmap_activity)<-activity
colnames(heatmap_activity)<-c(1:5)

heatmap(heatmap_activity, Rowv = NA, Colv = NA, main = 'STEM students perception of bioinformatics acti
```

EM students perception of bioinformatics activities