Teste_DTI

January 8, 2020

1 Analise da base de dados "wiki4HE"

1.0.1 A Utilização da Wikipedia como Ferramenta de Ensino

A base de dados "wiki4HE", é fruto de uma pesquisa enviada a professores de duas universidades espanholas entre 2012 e 2013: Universitat Oberta da Catalunya (UOC) e Universitat Pompeu Fabra (UPF). Ela contém 913 respostas (linhas) e 53 atributos (colunas).

A pesquisa foi organizada em duas partes. A primeira parte teve como objetivo coletar dados demográficos como: sexo, idade, área de especialização, doutorado, anos de experiência em ensino universitário, nível acadêmico e associação registrada na Wikipédia.

A segunda parte da pesquisa teve como objetivo reunir informações sobre os diferentes aspectos da Wikipedia no que diz respeito ao ensino superior e às opiniões dos professores. Essas perguntas tiveram que ser respondidas através de uma escala Likert de 5 pontos. Essa escala se refere ao nível de concordância ou discordância com uma afirmação (1 = "Discordo totalmente" e 5 = "Concordo totalmente").

Mais informações sobre a base de dados estão disponíveis no link: http://archive.ics.uci.edu/ml/datasets/wiki4HE.

1.0.2 Linguagem de programação utilizada

A linguagem de programação utilizada nas análises será o R (https://www.r-project.org/). O R é um ambiente de software livre para modelagens e gráficos estatísticos.

Todas as analises contidas neste documentos vão estar disponíveis no meu github. Link : https://github.com/MarinaAmorim/Analise-dos-dados-wiki4HE

Pacotes necessários para as analises

```
In [ ]: if (!require(dplyr)) install.packages('dplyr');library(dplyr)
        library(ggplot2) # Plotting
        library(knitr) # kable
        #library(GGally) # ggpairs plot
        library(ISLR) # Source of Data
        library(MASS) # Some Classification Models (LDA, QDA)
        library(class) #KNN
        library(caret) # Showing Confusion Matrix Data
        library(purrr) # Organizing
        library(tidyr) # Organize/tidy data
        #library(reshape) # Melt data for plotting
```

```
library(ape) # Trees
tableCounter = 0
figCounter = 0
#if (!require(SciencesPo)) install.packages('SciencesPo');library(SciencesPo)
```

Leitura da Base de dados Os dados estão em um arquivo excel (.csv), com o separador de ";" e os dados ausentes ou informações faltantes estão listados como "?".

```
In [42]: dados = read.csv("wiki4HE.csv", na.strings = "?", sep = ';') # leitura dos dados
```

```
In [7]: dim(dados) # dimensão dos dados ( linhas vs Colunas)
```

1. 913 2. 53

```
In [8]: head(dados) # as 6 primeiras linhas da base de dados
```

A data.frame: 6 Œ 53

AGE <int>	GENDER <int>	DOMAIN <int>	PhD <int>	YEARSEXP <int>	UNIVERSITY <int>	UOC_POSITIO <int>
40	0	2	1	14	1	2
42	0	5	1	18	1	2
37	0	4	1	13	1	3
40	0	4	0	13	1	3
51	0	6	0	8	1	3
47	0	4	0	17	1	3

Definindo os tipos de cada variável Dentre as várias informações observadas, devemos definir o tipo correto de cada variável para que ela receba o tratamento adequado na hora das analises. Uma variável pode ser numérica, categórica, lógica, etc.

```
In [44]: dados$GENDER <- as.factor( dados$GENDER )
         #table(dados$DOMAIN, useNA = "always")
         dados$DOMAIN = ifelse(dados$DOMAIN==6, NA,dados$DOMAIN)
         dados$DOMAIN <- as.factor( dados$DOMAIN )
         dados$PhD <- as.factor( dados$PhD )
         dados$YEARSEXP <- as.numeric( dados$YEARSEXP )
         dados$UNIVERSITY <- as.factor( dados$UNIVERSITY )
         dados$UOC_POSITION <- as.factor( dados$UOC_POSITION )
         dados$OTHER_POSITION <- as.factor( dados$OTHER_POSITION )
         dados$OTHERSTATUS <- as.factor( dados$OTHERSTATUS )
         dados$USERWIKI <- as.factor( dados$USERWIKI )
         dados[,11:53] <- lapply( dados[ , 11:53 ],  factor ) # restante das variáveis ( colun
```

```
In [173]: # Definindo o nome das categorias
          levels( dados$GENDER ) <- c( "Masculino", "Feminino" ) # sexo
          levels( dados$PhD ) <- c("Não","Sim" ) # phd
          levels( dados$UNIVERSITY ) <- c("UOC","UPF" ) # 1 = UOC; 2 = UPF
          levels( dados$USERWIKI ) <- c("Não","Sim" )#0=No; 1=Yes
          levels( dados$OTHER_POSITION ) <- c("Professor", "Associate", "NA" )#1=Professor; 2=
          levels( dados$DOMAIN ) <- c("Arts & Humanities","Sciences","Health Sciences",
```

```
                                        "Engineering & Architecture", "Law & Politics")
#1=Arts & Humanities; 2=Sciences; 3=Health Sciences; 4=Engineering & Architecture; 5=
levels( dados$UOC_POSITION ) <- c("Professor","Associate", "Assistant", "Lecturer",
                                  "Adjunct", "NA")
# 1=Professor; 2=Associate; 3=Assistant; 4=Lecturer; 5=Instructor; 6=Adjunct
```

Dados ausentes Sempre que formos analisar uma base de dados, devemos observar qual a proporção das informações faltantes, isso pode indicar algum problema na coleta ou até viesar os seus resultados. Como o nosso caso é um teste e não tem um objetivo específico, vamos apenas listar em quais variáveis tem essas informações faltantes, os chamados "NA's" e quantas são.

```
In [12]: #apply( dados, 2, anyNA )
```

```
In [61]: summary(dados)
```

```
      AGE              GENDER                                DOMAIN     PhD
 Min.   :23.00   Masculino:525   Arts & Humanities         :183   Não:489
 1st Qu.:36.00   Feminino :388   Sciences                  : 56   Sim:424
 Median :42.00                   Health Sciences           : 73
 Mean   :42.25                   Engineering & Architecture:137
 3rd Qu.:47.00                   Law & Politics            :101
 Max.   :69.00                   NA's                      :363

    YEARSEXP     UNIVERSITY    UOC_POSITION   OTHER_POSITION  OTHERSTATUS
 Min.   : 0.00   UOC:800    Adjunct  :659   Professor:268    2      :130
 1st Qu.: 5.00   UPF:113    Associate: 68   Associate:384    7      :107
 Median :10.00              Assistant: 50   NA       :  0    6      : 41
 Mean   :10.87              Lecturer : 18   NA's     :261    4      : 36
 3rd Qu.:15.00              Professor:  3                    3      : 24
 Max.   :43.00              (Other)  :  2                    (Other): 35
 NA's   :23                 NA's     :113                    NA's   :540
 USERWIKI     PU1         PU2         PU3         PEU1        PEU2        PEU3
 Não :784   1   : 35   1   : 33   1   : 20   1   :  3   1   :  3   1   : 17
 Sim :125   2   :216   2   :205   2   :151   2   : 21   2   : 35   2   : 97
 NA's:  4   3   :330   3   :339   3   :312   3   : 91   3   :166   3   :355
            4   :239   4   :244   4   :250   4   :328   4   :409   4   :250
            5   : 86   5   : 81   5   :175   5   :466   5   :286   5   : 97
            NA's:  7   NA's: 11   NA's:  5   NA's:  4   NA's: 14   NA's: 97

   ENJ1        ENJ2        Qu1         Qu2         Qu3         Qu4         Qu5
 1   : 19   1   :  4   1   : 24   1   : 11   1   : 28   1   : 39   1   : 54
 2   : 72   2   : 66   2   :163   2   :114   2   :236   2   :196   2   :183
 3   :207   3   :224   3   :371   3   :341   3   :389   3   :298   3   :366
 4   :386   4   :394   4   :308   4   :357   4   :215   4   :230   4   :234
 5   :222   5   :208   5   : 40   5   : 80   5   : 30   5   :128   5   : 47
 NA's:  7   NA's: 17   NA's:  7   NA's: 10   NA's: 15   NA's: 22   NA's: 29

   Vis1        Vis2        Vis3        Im1         Im2         Im3         SA1
```

```
1   : 49   1   : 25   1   :406   1   :139   1   : 34   1   : 73   1   :  5
2   :190   2   :145   2   :222   2   :335   2   :181   2   :218   2   : 26
3   :396   3   :422   3   :161   3   :294   3   :273   3   :342   3   :142
4   :170   4   :158   4   : 79   4   : 98   4   :298   4   :178   4   :348
5   : 36   5   : 46   5   : 37   5   : 25   5   :107   5   : 45   5   :381
NA's: 72   NA's:117   NA's:  8   NA's: 22   NA's: 20   NA's: 57   NA's: 11

  SA2        SA3        Use1       Use2       Use3       Use4       Use5
1   : 13   1   :  4   1   :329   1   :471   1   :194   1   :212   1   : 55
2   : 50   2   : 21   2   :272   2   :207   2   :226   2   :232   2   :152
3   :140   3   : 96   3   :191   3   :137   3   :238   3   :237   3   :278
4   :302   4   :285   4   : 79   4   : 60   4   :184   4   :159   4   :290
5   :396   5   :496   5   : 28   5   : 21   5   : 62   5   : 50   5   :123
NA's: 12   NA's: 11   NA's: 14   NA's: 17   NA's:  9   NA's: 23   NA's: 15

  Pf1        Pf2        Pf3        JR1        JR2        BI1        BI2
1   :340   1   :167   1   :257   1   : 32   1   :105   1   : 80   1   : 84
2   :218   2   :223   2   :199   2   : 98   2   :146   2   :206   2   :207
3   :169   3   :215   3   :212   3   :203   3   :277   3   :344   3   :298
4   :107   4   :173   4   :153   4   :325   4   :215   4   :178   4   :196
5   : 68   5   :129   5   : 78   5   :228   5   :117   5   : 73   5   : 85
NA's: 11   NA's:  6   NA's: 14   NA's: 27   NA's: 53   NA's: 32   NA's: 43

  Inc1       Inc2       Inc3       Inc4       Exp1       Exp2       Exp3
1   : 47   1   : 55   1   : 55   1   : 67   1   :102   1   : 46   1   : 37
2   : 62   2   :113   2   :127   2   :102   2   :232   2   :135   2   :126
3   :209   3   :260   3   :257   3   :249   3   :223   3   :222   3   :174
4   :309   4   :272   4   :250   4   :243   4   :249   4   :327   4   :340
5   :251   5   :178   5   :187   5   :210   5   : 94   5   :172   5   :223
NA's: 35   NA's: 35   NA's: 37   NA's: 42   NA's: 13   NA's: 11   NA's: 13

  Exp4       Exp5
1   :580   1   :292
2   :176   2   :210
3   : 91   3   :153
4   : 37   4   :158
5   : 15   5   : 87
NA's: 14   NA's: 13
```

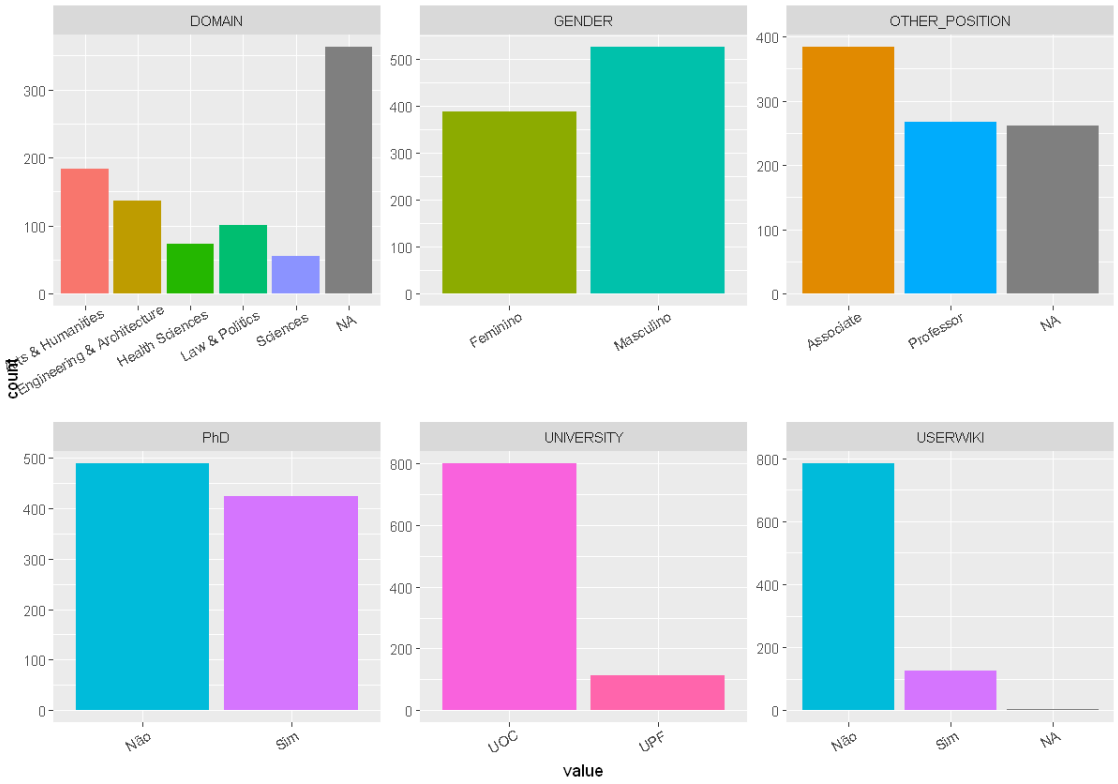
Apenas as variáveis AGE, GENDER, PhD, UNIVERSITY não possuem informações faltantes. Mas, para os fins pretendidos esses NA's não vão ser um problema.

2 Análise Exploratória

A seguir vamos apresentar alguns gráficos e tabelas que vão sumarizar as informações contida em nossa base de dados.

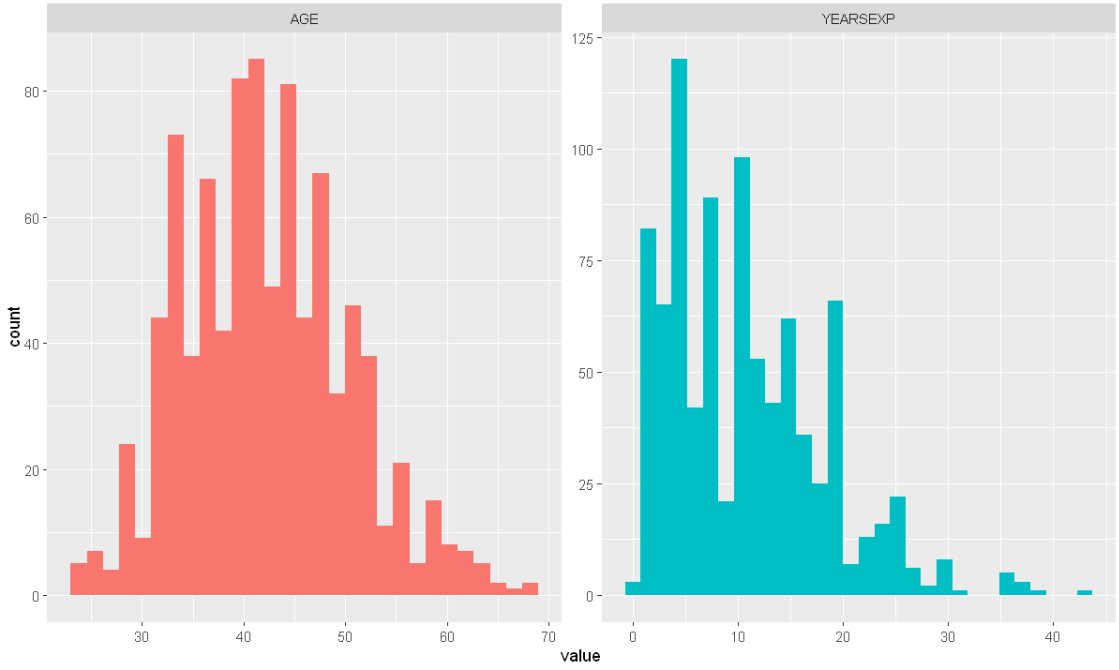
In [69]:
```
options(repr.plot.width=10, repr.plot.height=7)
wikifactor = dados # Create separate copy for changing survey items to ordered/factor
wikifactor[,11:ncol(dados)]=lapply(wikifactor[,11:ncol(dados)], ordered)
wikifactor[,c("GENDER","PhD","UNIVERSITY","USERWIKI", "OTHER_POSITION","DOMAIN")] %>%
  keep(is.factor) %>%
  gather() %>%
  ggplot(aes(value, fill=value)) +
  facet_wrap(~ key, scales = "free") +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 30, hjust = 0.85),legend.position="none")
```

```
Warning message:
"attributes are not identical across measure variables;
they will be dropped"
```



In [66]:
```
options(repr.plot.width=10, repr.plot.height=6)
wikifactor %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value,fill=key)) +
  facet_wrap(~ key, scales = "free") +
  geom_histogram(bins=sqrt(nrow(dados))) +
  theme(legend.position="none")
```

```
Warning message:
"Removed 23 rows containing non-finite values (stat_bin)."
```



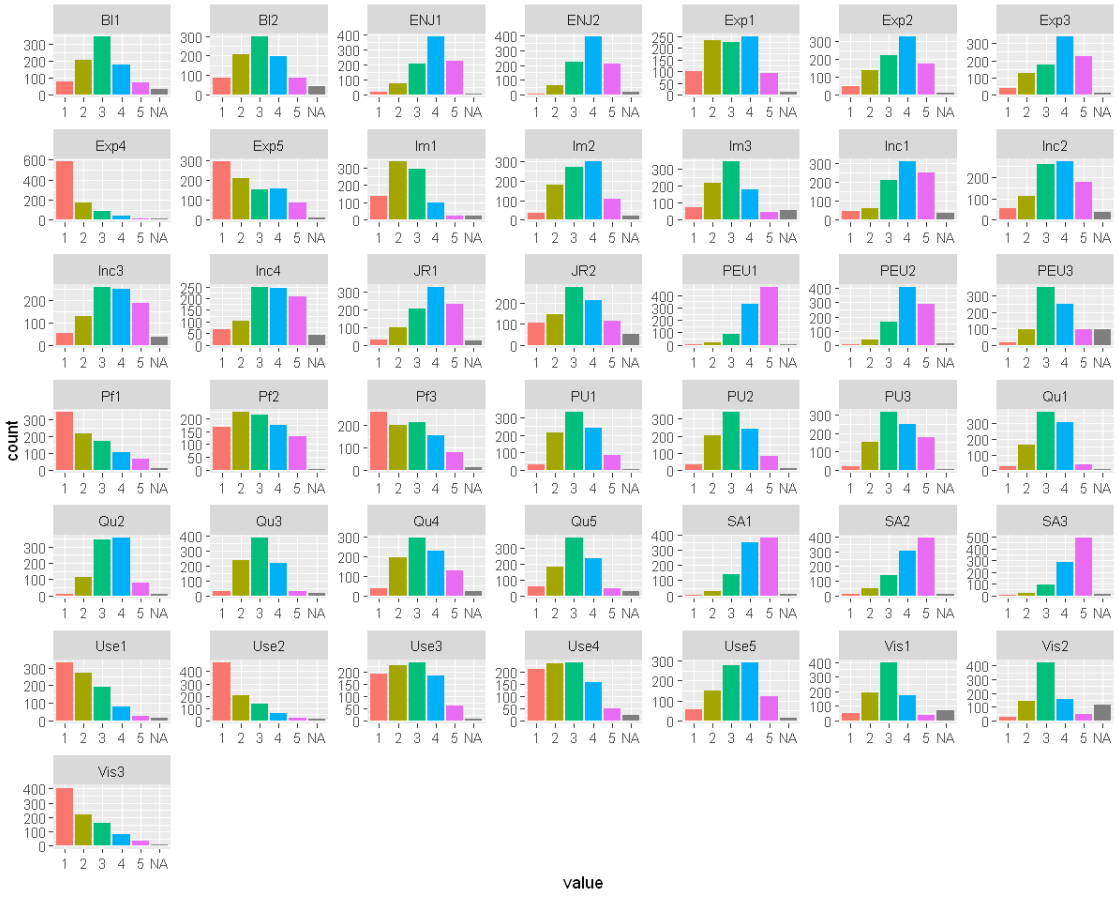
Tabelas

```
In [60]: dados %>% dplyr::select(c( GENDER,PhD,UNIVERSITY,USERWIKI, OTHER_POSITION,DOMAIN, AGE
```

```
      GENDER     PhD      UNIVERSITY USERWIKI     OTHER_POSITION
 Masculino:525   Não:489   UOC:800    Não :784   Professor:268
 Feminino :388   Sim:424   UPF:113    Sim :125   Associate:384
                                      NA's:  4   NA       :  0
                                                 NA's     :261


                        DOMAIN         AGE           YEARSEXP
 Arts & Humanities         :183   Min.   :23.00   Min.   : 0.00
 Sciences                  : 56   1st Qu.:36.00   1st Qu.: 5.00
 Health Sciences           : 73   Median :42.00   Median :10.00
 Engineering & Architecture:137   Mean   :42.25   Mean   :10.87
 Law & Politics            :101   3rd Qu.:47.00   3rd Qu.:15.00
 NA's                      :363   Max.   :69.00   Max.   :43.00
                                                  NA's   :23
```

Variáveis que estão na escala Likert de 5 pontos

```
In [70]: options(repr.plot.width=10, repr.plot.height=8)
         wikifactor[11:ncol(wikifactor)] %>%
           keep(is.factor) %>%
           gather() %>%
           ggplot(aes(value,fill=value)) +
           facet_wrap(~ key, scales = "free") +
           geom_bar()+
           theme(legend.position="none")
```



2.1 Análise das respostas entre os diferentes grupos de usuários

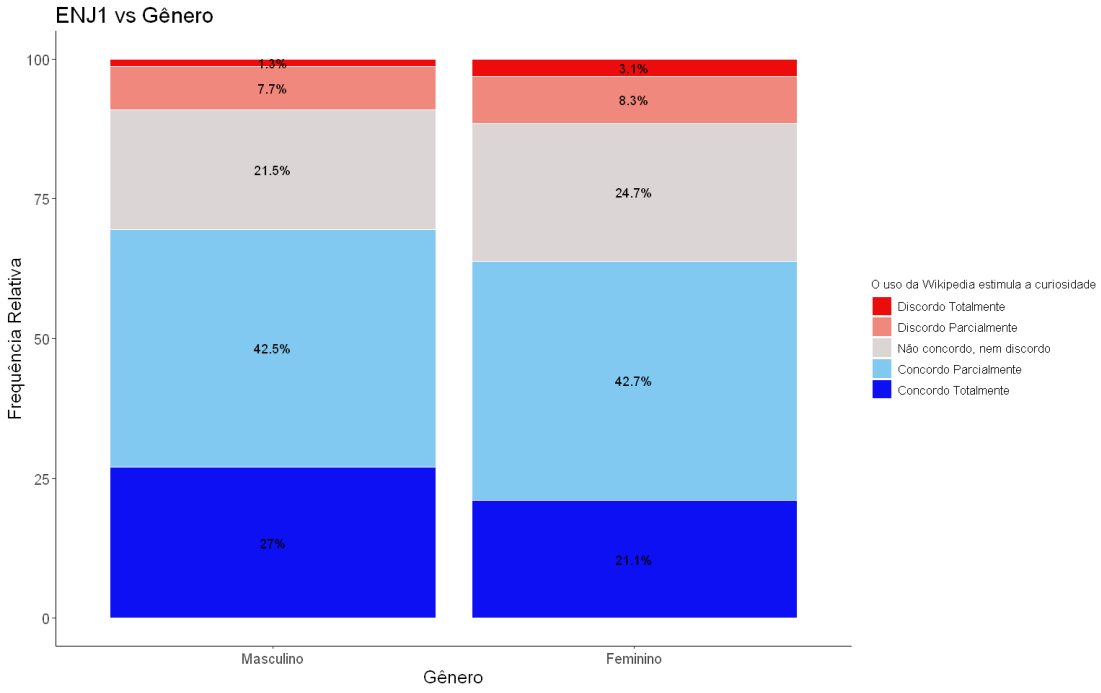
Nesta parte do trabalho, vamos analisar as respostas de diferentes grupos de usuários para os itens da categoria "Perceived Enjoyment": ENJ1 e ENJ2.

* ENJ1: O uso da Wikipedia estimula a curiosidade
* ENJ2: O uso da Wikipedia é divertido

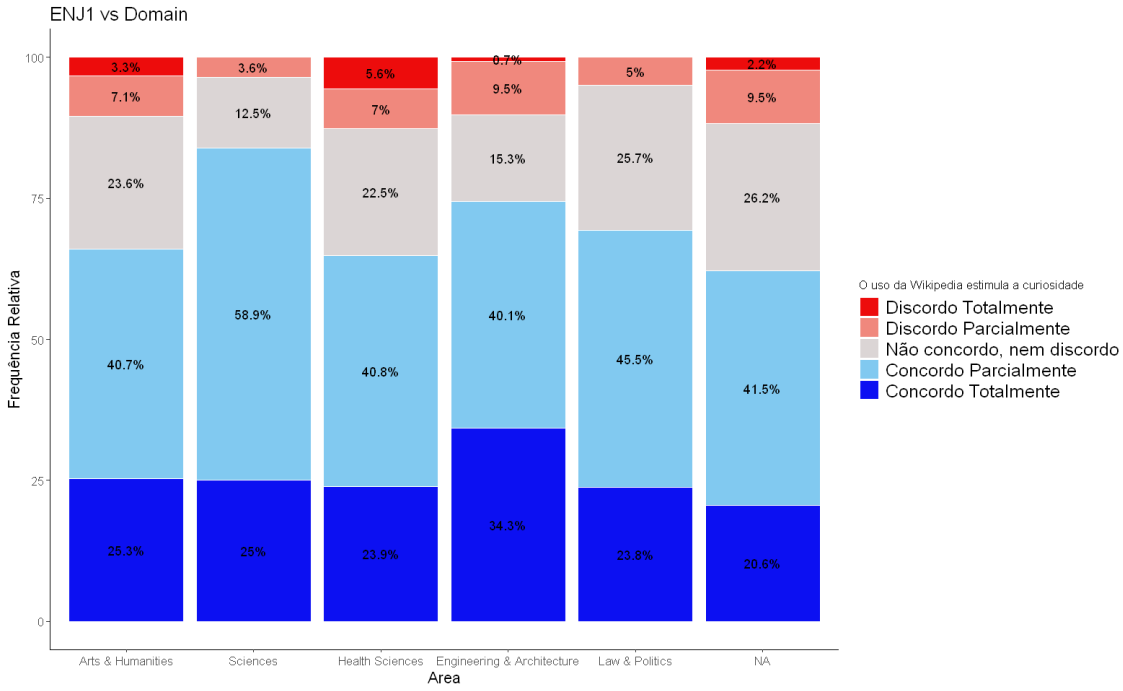
```
In [113]: # definindo as categórias
          levels( dados$ENJ1 ) = c( "Discordo Totalmente", "Discordo Parcialmente",
                                    "Não concordo, nem discordo",
                                    "Concordo Parcialmente", "Concordo Totalmente")


          levels( dados$ENJ2 ) = c( "Discordo Totalmente", "Discordo Parcialmente",
                                    "Não concordo, nem discordo",
                                    "Concordo Parcialmente", "Concordo Totalmente")
```

2.1.1 O uso da Wikipédia estimula a criatividade?

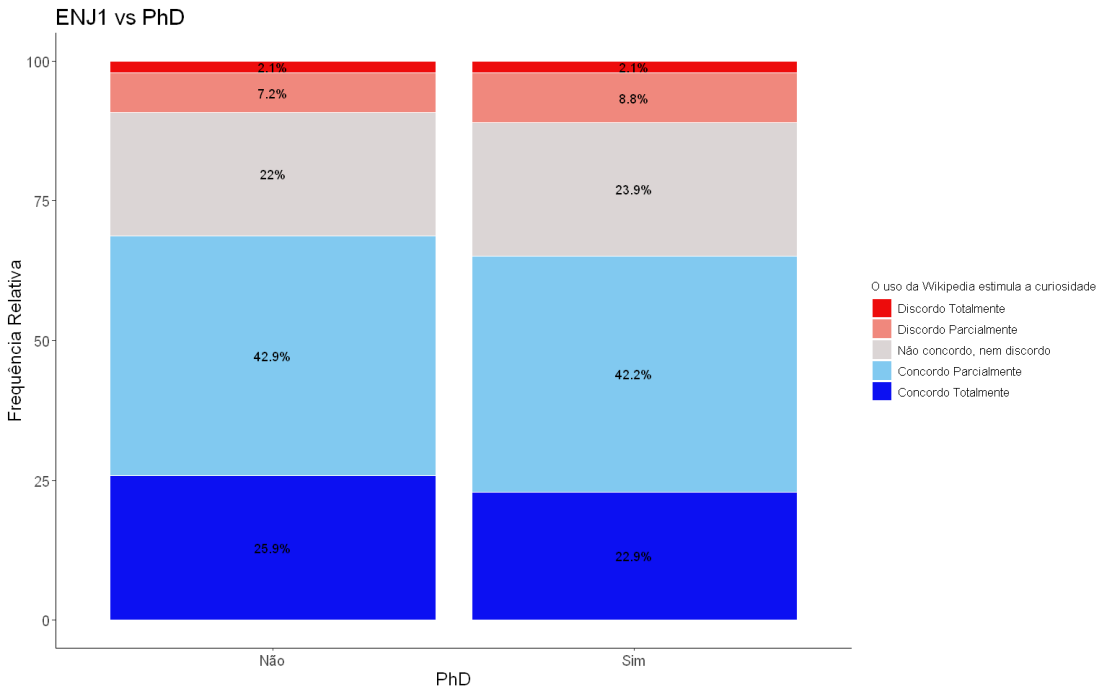
```
In [171]: mycols = c("#ed0c0c","#f0887d", "#dbd5d5", "#81c9f0","#0c10f2" )
          ### ENJ1 vs gender ==========

          ENJ1_gender <- dados %>%
            drop_na(ENJ1) %>% # drop 7 observations
            group_by(GENDER, ENJ1) %>%
            summarise(n = n()) %>%
            mutate( freq = (n / sum(n))*100  )  # table of frequency and relative frequency to

          options(repr.plot.width=13, repr.plot.height=8)
          ggplot( ENJ1_gender, aes(x = GENDER, y = freq, fill = ENJ1, label = paste0(round(freq
            geom_bar(stat = "identity", color = "white" ) +
              geom_text(position = position_stack(vjust = 0.5))+
          ggtitle("ENJ1 vs Gênero")+
            labs(fill = "O uso da Wikipedia estimula a curiosidade", y = "Frequência Relativa"
            scale_fill_manual(values = mycols) +
            theme_classic() +
            theme(legend.title = element_text(size = 10),
                  legend.text = element_text(size = 10),
                  text = element_text(size=15))
```



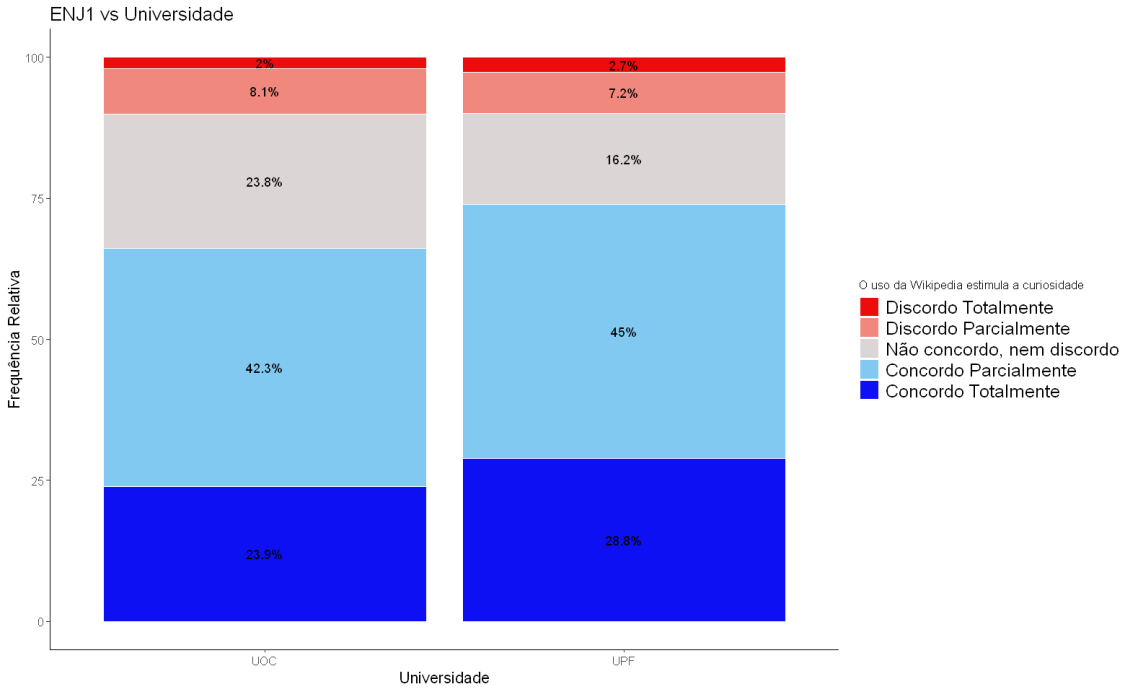
In [170]:
```
### ENJ1 vs Area ==========

ENJ1_domain <- dados %>%
  drop_na(ENJ1) %>% # drop 7 observations
  group_by(DOMAIN, ENJ1) %>%
  summarise(n = n()) %>%
  mutate( freq = (n / sum(n))*100  )  # table of frequency and relative frequency to

options(repr.plot.width=13, repr.plot.height=8)
ggplot( ENJ1_domain, aes(x = DOMAIN, y = freq, fill = ENJ1, label = paste0(round(freq
  geom_bar(stat = "identity", color = "white" ) +
ggtitle("ENJ1 vs Domain")+
    geom_text(position = position_stack(vjust = 0.5))+
  labs(fill = "O uso da Wikipedia estimula a curiosidade", y = "Frequência Relativa"
  scale_fill_manual(values = mycols) +
  theme_classic() +
  theme(legend.title = element_text(size = 10),
        legend.text = element_text(size = 15),
        text = element_text(size=13))
```



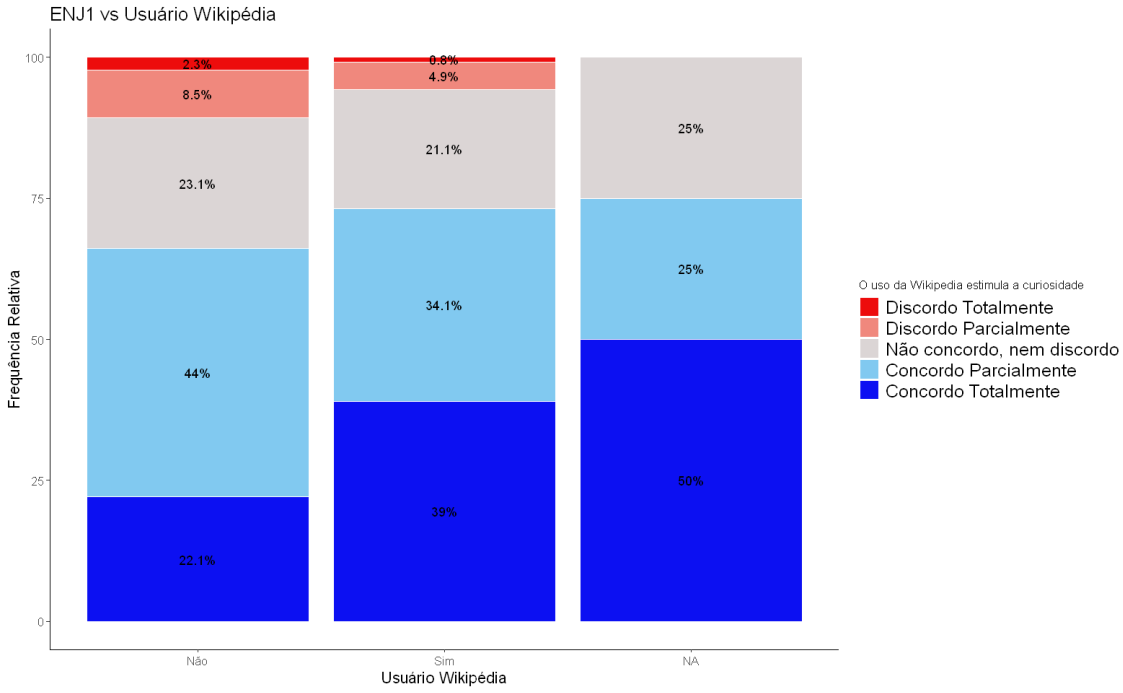
In [169]:
```
### ENJ1 vs Phd ==========

ENJ1_phd <- dados %>%
  drop_na(ENJ1) %>% # drop 7 observations
  group_by(PhD, ENJ1) %>%
  summarise(n = n()) %>%
  mutate( freq = (n / sum(n))*100  )  # table of frequency and relative frequency to

options(repr.plot.width=13, repr.plot.height=8)
ggplot( ENJ1_phd, aes(x = PhD, y = freq, fill = ENJ1, label = paste0(round(freq, 1),'
  geom_bar(stat = "identity", color = "white" ) +
ggtitle("ENJ1 vs PhD")+
    geom_text(position = position_stack(vjust = 0.5))+
  labs(fill = "O uso da Wikipedia estimula a curiosidade", y = "Frequência Relativa"
  scale_fill_manual(values = mycols) +
  theme_classic() +
  theme(legend.title = element_text(size = 10),
        legend.text = element_text(size = 10),
        text = element_text(size=15))
```



In [168]:
```
### ENJ1 vs Universidade ==========

ENJ1_uni <- dados %>%
  drop_na(ENJ1) %>% # drop 7 observations
  group_by(UNIVERSITY, ENJ1) %>%
  summarise(n = n()) %>%
  mutate( freq = (n / sum(n))*100  )  # table of frequency and relative frequency to

options(repr.plot.width=13, repr.plot.height=8)
ggplot( ENJ1_uni, aes(x = UNIVERSITY, y = freq, fill = ENJ1, label = paste0(round(fr
  geom_bar(stat = "identity", color = "white" ) +
ggtitle("ENJ1 vs Universidade")+
geom_text(position = position_stack(vjust = 0.5))+
  labs(fill = "O uso da Wikipedia estimula a curiosidade", y = "Frequência Relativa"
  scale_fill_manual(values = mycols) +
  theme_classic() +
  theme(legend.title = element_text(size = 10),
        legend.text = element_text(size = 15),
        text = element_text(size=13))
```



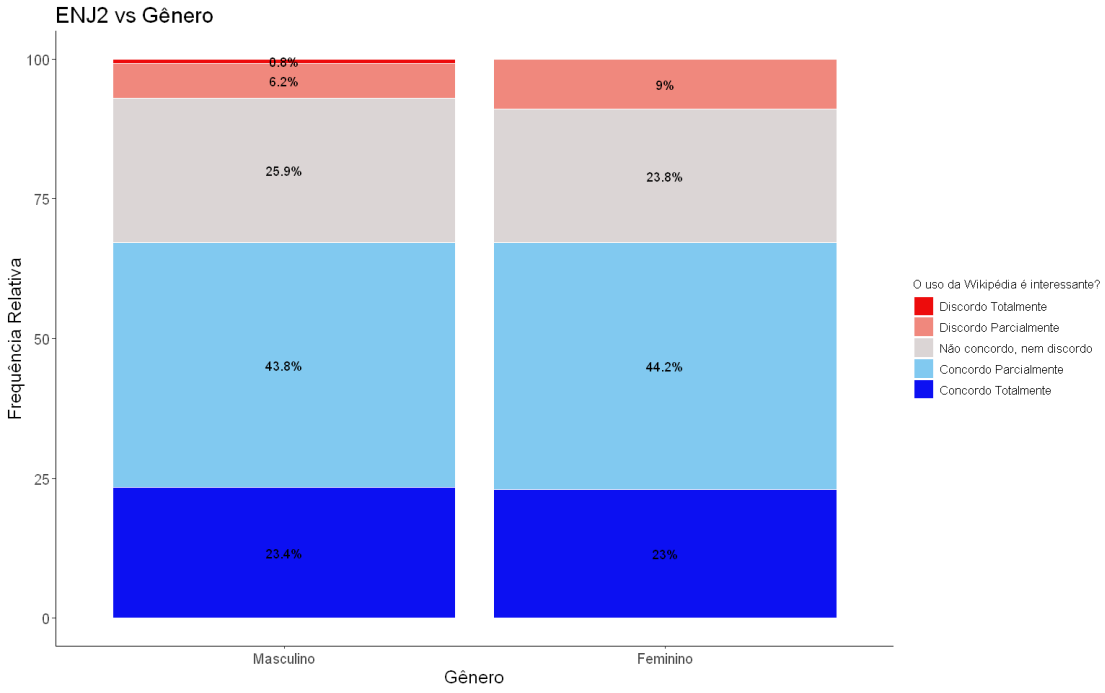
In [167]:
```
### ENJ1 vs Usuário Wikipédia ==========

ENJ1_user <- dados %>%
  drop_na(ENJ1) %>% # drop 7 observations
  group_by(USERWIKI, ENJ1) %>%
  summarise(n = n()) %>%
  mutate( freq = (n / sum(n))*100  )  # table of frequency and relative frequency to

options(repr.plot.width=13, repr.plot.height=8)
ggplot( ENJ1_user, aes(x = USERWIKI, y = freq, fill = ENJ1, label = paste0(round(freq
  geom_bar(stat = "identity", color = "white" ) +
ggtitle("ENJ1 vs Usuário Wikipédia")+
    geom_text(position = position_stack(vjust = 0.5))+
  labs(fill = "O uso da Wikipedia estimula a curiosidade", y = "Frequência Relativa"
  scale_fill_manual(values = mycols) +
  theme_classic() +
  theme(legend.title = element_text(size = 10),
        legend.text = element_text(size = 15),
        text = element_text(size=13))
```



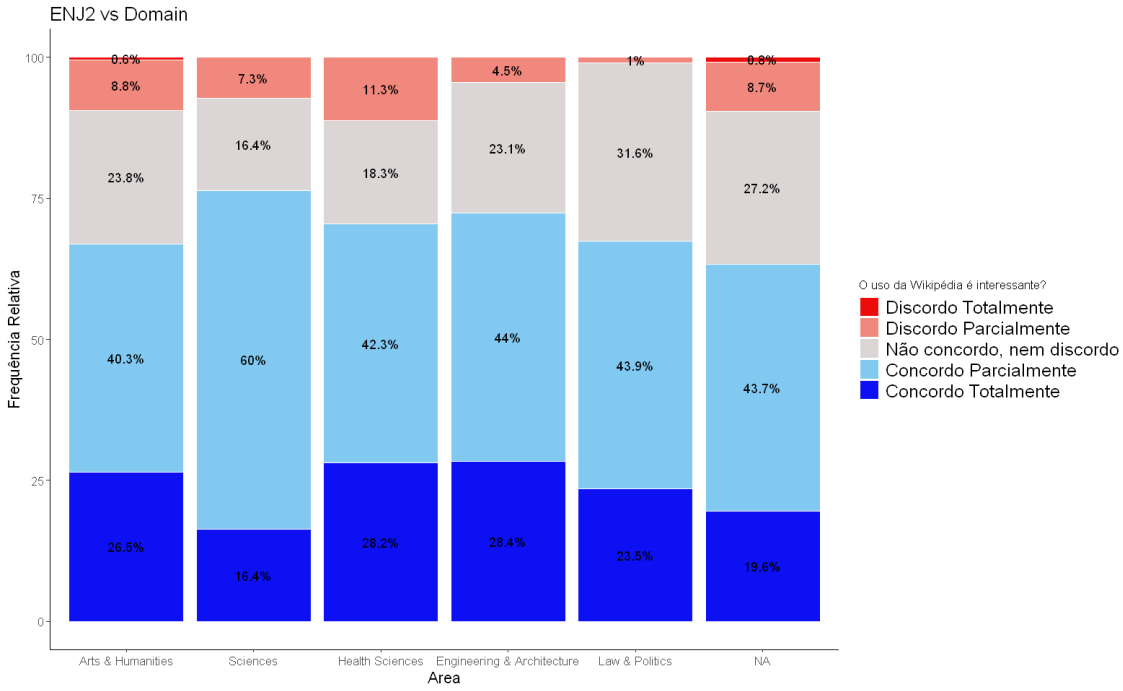
Quando analisamos a variável ENJ1 (O uso da Wikipedia estimula a curiosidade) comparada com sexo, area, PhD, universidade e se utiliza a wikipédia, os percentuais são muito parecidos entre as categórias. Apenas a variável que leva em consideração o fato do professor utilizar ou não a wikipédia que mostra uma leve diferença nas categórias, mostrando que indivíduos que usam a ferramenta tende a concordar com o fato que ela estimula a criatividade.

2.1.2 O uso da Wikipédia é interessante?

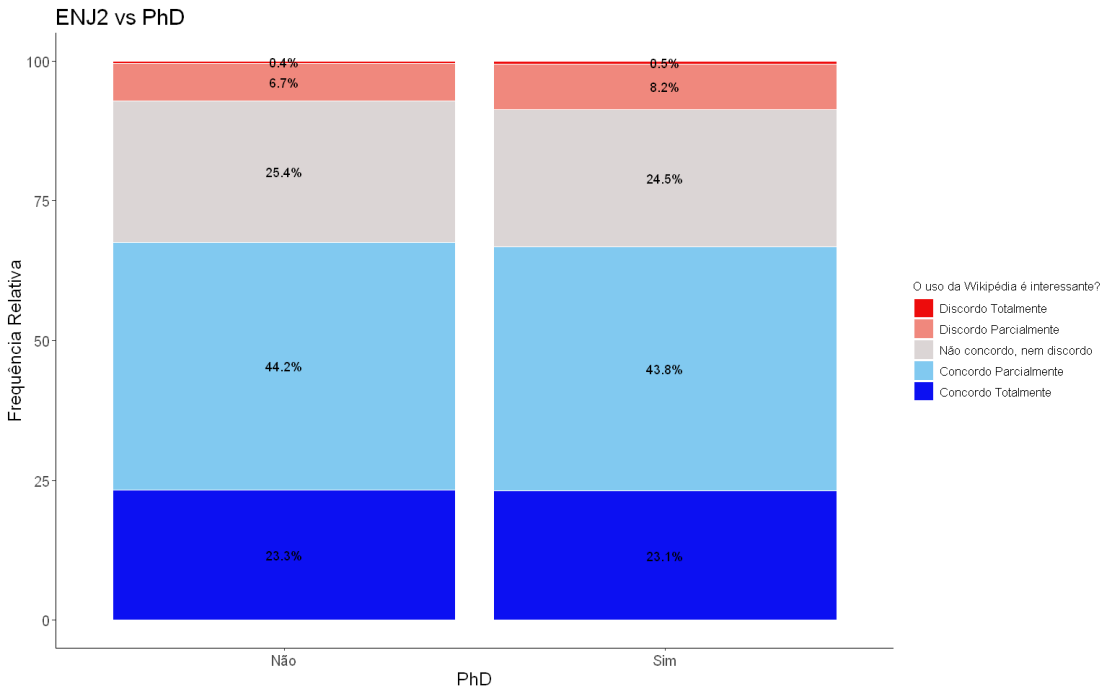
```
In [166]: ### ENJ2 vs Gênero ==========
          ENJ2_gender <- dados %>%
            drop_na(ENJ2) %>% # drop 7 observations
            group_by(GENDER, ENJ2) %>%
            summarise(n = n()) %>%
            mutate( freq = (n / sum(n))*100  )  # table of frequency and relative frequency to

          options(repr.plot.width=13, repr.plot.height=8)
          ggplot( ENJ2_gender, aes(x = GENDER, y = freq, fill = ENJ2, label = paste0(round(freq
            geom_bar(stat = "identity", color = "white" ) +
          ggtitle("ENJ2 vs Gênero")+
              geom_text(position = position_stack(vjust = 0.5))+
            labs(fill = "O uso da Wikipédia é interessante?", y = "Frequência Relativa", x = "G
            scale_fill_manual(values = mycols) +
            theme_classic() +
            theme(legend.title = element_text(size = 10),
                  legend.text = element_text(size = 10),
                  text = element_text(size=15))
```



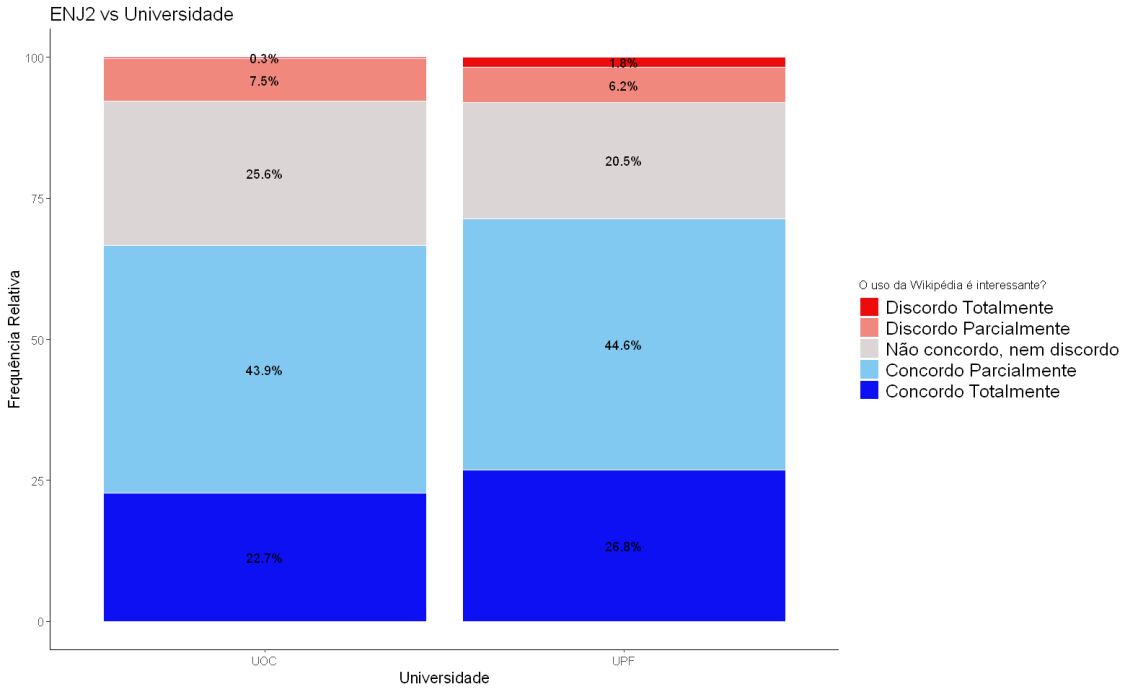
In [165]:
```
### ENJ2 vs Area ==========

ENJ2_domain <- dados %>%
  drop_na(ENJ2) %>% # drop 7 observations
  group_by(DOMAIN, ENJ2) %>%
  summarise(n = n()) %>%
  mutate( freq = (n / sum(n))*100  )  # table of frequency and relative frequency to

options(repr.plot.width=13, repr.plot.height=8)
ggplot( ENJ2_domain, aes(x = DOMAIN, y = freq, fill = ENJ2, label = paste0(round(freq
  geom_bar(stat = "identity", color = "white" ) +
ggtitle("ENJ2 vs Domain")+
    geom_text(position = position_stack(vjust = 0.5))+
  labs(fill = "O uso da Wikipédia é interessante?", y = "Frequência Relativa", x = "A
  scale_fill_manual(values = mycols) +
  theme_classic() +
  theme(legend.title = element_text(size = 10),
        legend.text = element_text(size = 15),
        text = element_text(size=13))
```



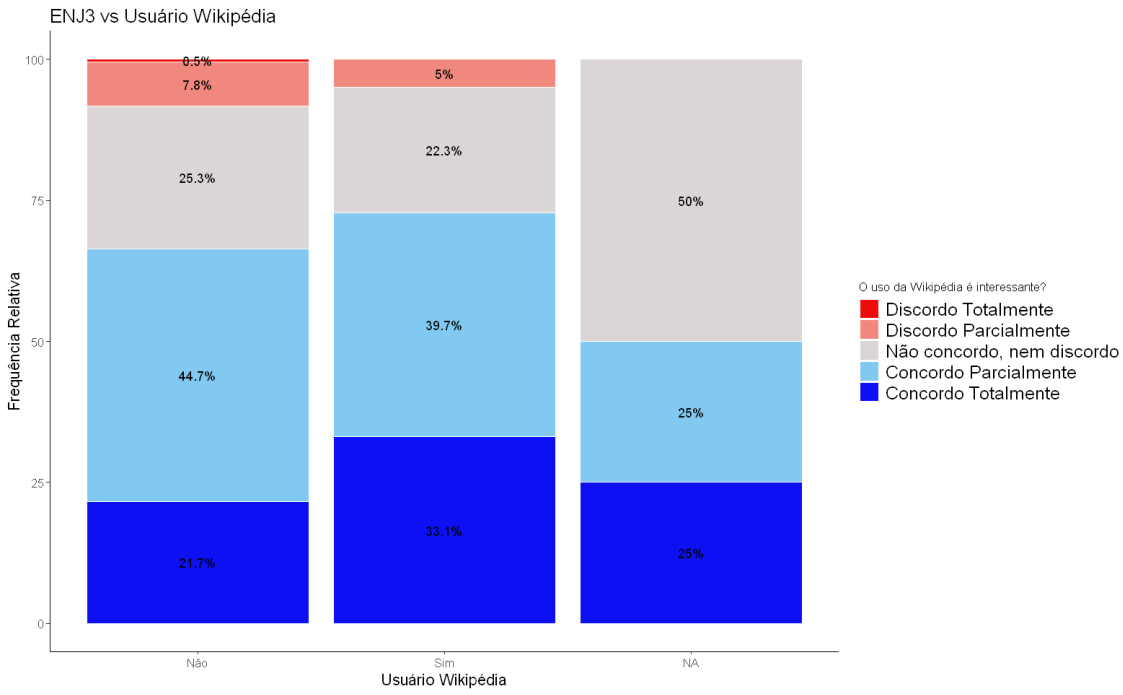
In [164]:
```
### ENJ2 vs Phd ==========

ENJ2_phd <- dados %>%
  drop_na(ENJ2) %>% # drop 7 observations
  group_by(PhD, ENJ2) %>%
  summarise(n = n()) %>%
  mutate( freq = (n / sum(n))*100  )  # table of frequency and relative frequency to

options(repr.plot.width=13, repr.plot.height=8)
ggplot( ENJ2_phd, aes(x = PhD, y = freq, fill = ENJ2, label = paste0(round(freq, 1),'
  geom_bar(stat = "identity", color = "white" ) +
ggtitle("ENJ2 vs PhD")+
    geom_text(position = position_stack(vjust = 0.5))+
  labs(fill = "O uso da Wikipédia é interessante?", y = "Frequência Relativa", x = "I
  scale_fill_manual(values = mycols) +
  theme_classic() +
  theme(legend.title = element_text(size = 10),
        legend.text = element_text(size = 10),
        text = element_text(size=15))
```



In [163]:
```
### ENJ2 vs Universidade ==========

ENJ2_uni <- dados %>%
  drop_na(ENJ2) %>% # drop 7 observations
  group_by(UNIVERSITY, ENJ2) %>%
  summarise(n = n()) %>%
  mutate( freq = (n / sum(n))*100  )  # table of frequency and relative frequency to

options(repr.plot.width=13, repr.plot.height=8)
ggplot( ENJ2_uni, aes(x = UNIVERSITY, y = freq, fill = ENJ2, label = paste0(round(fre
  geom_bar(stat = "identity", color = "white" ) +
    geom_text(position = position_stack(vjust = 0.5))+
ggtitle("ENJ2 vs Universidade")+
  labs(fill = "O uso da Wikipédia é interessante?", y = "Frequência Relativa", x = "U
  scale_fill_manual(values = mycols) +
  theme_classic() +
  theme(legend.title = element_text(size = 10),
        legend.text = element_text(size = 15),
        text = element_text(size=13))
```



In [162]:
```
### ENJ2 vs Usuário Wikipédia ==========

ENJ2_user <- dados %>%
  drop_na(ENJ2) %>% # drop 7 observations
  group_by(USERWIKI, ENJ2) %>%
  summarise(n = n()) %>%
  mutate( freq = (n / sum(n))*100  )  # table of frequency and relative frequency to

options(repr.plot.width=13, repr.plot.height=8)
ggplot( ENJ2_user, aes(x = USERWIKI, y = freq, fill = ENJ2, label = paste0(round(freq
  geom_bar(stat = "identity", color = "white" ) +
ggtitle("ENJ3 vs Usuário Wikipédia")+
    geom_text(position = position_stack(vjust = 0.5))+
  labs(fill = "O uso da Wikipédia é interessante?", y = "Frequência Relativa", x = "U
  scale_fill_manual(values = mycols) +
  theme_classic() +
  theme(legend.title = element_text(size = 10),
        legend.text = element_text(size = 15),
        text = element_text(size=13))
```



Já para a variável ENJ2 (O uso da Wikipedia é divertido) comparada com sexo, area, PhD, universidade e se utiliza a wikipédia, os percentuais são muito parecidos entre as categórias, os resultados são bem parecidos com os anteriores. Apenas a variável que leva em consideração o fato do professor utilizar ou não a wikipédia que mostra uma leve diferença nas categórias, mostrando que indivíduos que usam a ferramenta tende a concordar com o fato que ela é divertida.

2.2 Modelagem Estatística

2.2.1 Recomendações (USE3)

Em nossa base de dados temos a variável **USE3** (recomendo que meus alunos usem a Wikipédia), está variável está na escala likert, ou seja, de 1 a 5. Vamos criar um indicador para transformar a variável em binária. Para isso, vamos definir que professores que marcaram para a questão os valores 4 ou 5 recomendariam o uso e os que marcaram 1,2 e 3 não recomendariam. Essa regra de decisão é arbitrária e poderia ser feita de outra forma se preferir.

```
In [203]: dados$recomenda = ifelse(dados$Use3=="5"|dados$Use3=="4", "1", "0"  ) # 1 = sim e 0 =
          dados$recomenda = as.factor(dados$recomenda)
          levels(dados$recomenda)=c( "Não","Sim")
          table(dados$recomenda)
```

```
Não Sim
658 246
```

Agora que a variável de recomendação do professor a wikipédia é uma variável binária, podemos utilizar regressão logística. A regressão logística é uma técnica estatística que tem como objetivo produzir, a partir de um conjunto de observações, um modelo que permita a predição de valores tomados por uma variável categórica, binária, a partir de uma série de variáveis explicativas. Como variável explicativa vamos utilizar as questões que já discutimos neste relatório, são elas: Idade, Gênero, Universidade, Anos de Experiência, Area, Se tem ou não PhD e se utiliza a wikipédia.

Com a regressão logística podemos identificar quais variáveis são importantes para explicar o fato do dos professores recomendarem ou não o uso da wikipédia e com isso, conseguimos traçar os perfins dos professores que recomendam.

```
In [198]: mod = glm( recomenda ~ AGE + GENDER + DOMAIN + PhD + YEARSEXP + UNIVERSITY + USERWIKI
                     data = dados, family = binomial( link = "logit" ) )
          summary(mod)
```

```
Call:
glm(formula = recomenda ~ AGE + GENDER + DOMAIN + PhD + YEARSEXP +
    UNIVERSITY + USERWIKI, family = binomial(link = "logit"),
    data = dados)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8702  -0.8370  -0.5371   0.7738   2.5249

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -0.345436   0.703492  -0.491 0.623405
AGE                             -0.008976   0.016701  -0.537 0.590957
GENDERFeminino                  -0.959834   0.253262  -3.790 0.000151 ***
DOMAINSciences                   0.317543   0.366804   0.866 0.386653
DOMAINHealth Sciences           -0.655760   0.367540  -1.784 0.074393 .
DOMAINEngineering & Architecture -0.264385   0.288523  -0.916 0.359490
DOMAINLaw & Politics            -1.636542   0.401913  -4.072 4.66e-05 ***
PhDSim                          -0.160300   0.252218  -0.636 0.525061
YEARSEXP                         0.018646   0.019872   0.938 0.348102
UNIVERSITYUPF                    0.693490   0.315851   2.196 0.028119 *
USERWIKISim                      1.874434   0.292046   6.418 1.38e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 637.11  on 524  degrees of freedom
Residual deviance: 535.14  on 514  degrees of freedom
  (388 observations deleted due to missingness)
AIC: 557.14
```

```
Number of Fisher Scoring iterations: 5
```

Após ajustar o modelo, temos um teste de hipótese para cada variável explicativa, esse teste nos informa quais variáveis são estatisticamente significativas para explicar a recomendação da wikipédia para os alunos. Concluimos que, apenas as variáveis Gênero, Domain, Universidade e o fato de usar ou não a wikipédia ajuda a explicar o fato do professor recomendar ou não o uso para os alunos.