Avance 2

Grupo 6

21/1/2023

##Visualización de Datos

Luego de haber scrappeado las páginas que ofrecen empleos en el ámbito de desarrolladores a nivel nacional, se obtiene el archivo "final.csv" el cual es el data frame a usarse para responder a las preguntas del proyecto por medio de una visualización con el uso de R.

```
library(readr)
library(stringr)
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
datos_final <- read.csv("final.csv")
summary(datos_final)
```

```
##       num       titulo_empleo      descripcion_empleo etiquetas_empleo
##  Min.   :   0   Length:8556        Length:8556        Length:8556
##  1st Qu.:1041   Class :character   Class :character   Class :character
##  Median :3172   Mode  :character   Mode  :character   Mode  :character
##  Mean   :3308
##  3rd Qu.:5310
##  Max.   :7449
```

Filtramos datos por modalidad

Para realizar la filtración de modalidad, nos basamos en la descripción del empleo en el cual se obtiene información acerca del tipo de trabajo que se ofrece, en este caso nos importa la modalidad, para esto usamos la función str_detect para poder usar su subfunción "regex" con el fin de poder escribir una expresión regular, y además colocamos el valor de TRUE en ignore_case para que no importe si se encuentra escrita entre las mayúsculas o minúsculas. Hay que tener en cuenta que habrán descripciones en las que no se mencione la modalidad. Pero con estos datos se va a trabajar para responder algunas preguntas.

```
#PRESENCIAL
descripcion <- datos_final$descripcion_empleo
presencial <- str_detect(descripcion, regex("presencial|(\\s+|,|^)on site(\\s+|,|$)", ignore_case = TRU
desc_presencial <- descripcion[presencial]

#REMOTO
remota <- str_detect(descripcion, regex("remote|remota|remoto", ignore_case = TRUE))
desc_remota <- descripcion[remota]

#HÍBRIDO
hibrido <- str_detect(descripcion, regex("híbrido|híbrida|hybrid|hibrido|hibrida|semipresencial", ignor
desc_hibrido <- descripcion[hibrido]
```

Preguntas Alan

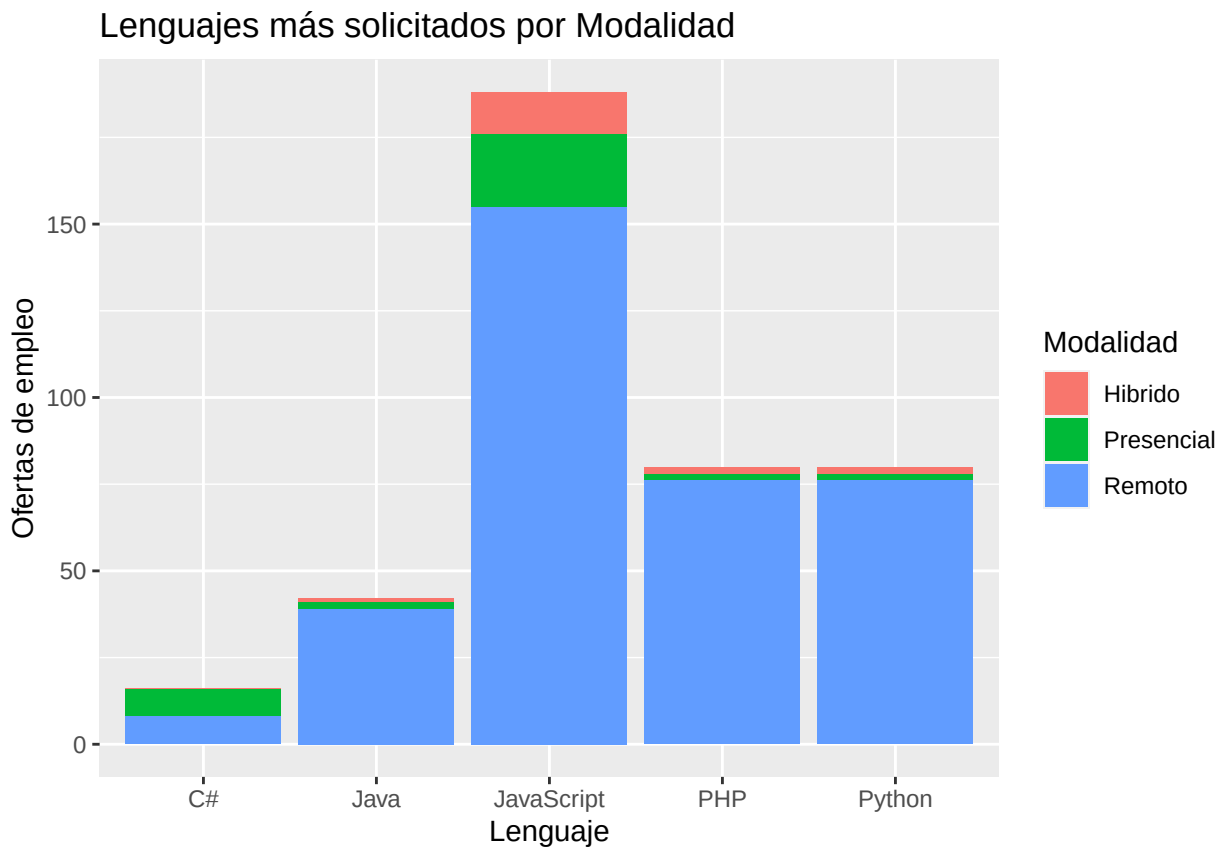
Pregunta 1 ¿Qué lenguajes de programación son los más solicitados por modalidad de trabajo?

```
#PRESENCIAL
#PHP
php_pres <- str_detect(desc_presencial, regex("php|laravel", ignore_case = TRUE))

#Javascript
javascript_pres <- str_detect(desc_presencial, regex("javascript|js|react", ignore_case = TRUE))



#python
python_pres <- str_detect(desc_presencial, regex("python|django", ignore_case = TRUE))

#c#
csharp_pres <- str_detect(desc_presencial, regex("c#", ignore_case = TRUE))


#java
java_pres <- str_detect(desc_presencial,"(\\s+|,|^)[Jj]ava(\\s+|,|$)")



lenguajes_pres <- c("PHP","JavaScript","Python","C#","Java")
valores_pres <- c(sum(python_pres == TRUE),sum(javascript_pres == TRUE),sum(python_pres == TRUE),
                  sum(csharp_pres==TRUE),sum(java_pres==TRUE))

#############DATA FRAME PRESENCIAL     ###################
df_pres <- data.frame("lenguajes"=lenguajes_pres,"valores"=valores_pres)
df_pres <- cbind(df_pres,Modalidad="Presencial")


#REMOTA ######################
#PHP
php_remo <- str_detect(desc_remota, regex("php|laravel", ignore_case = TRUE))
```

```r
#Javascript
javascript_remo <- str_detect(desc_remota, regex("javascript|js|react", ignore_case = TRUE))


#python
python_remo <- str_detect(desc_remota, regex("python|django", ignore_case = TRUE))

#c#
csharp_remo <- str_detect(desc_remota, regex("c#", ignore_case = TRUE))


#java
java_remo <- str_detect(desc_remota,"(\\s+|,|^)[Jj]ava(\\s+|,|$)")

######DATA FRAME ##################
lenguajes_remo <- c("PHP","JavaScript","Python","C#","Java")
valores_remo <- c(sum(python_remo == TRUE),sum(javascript_remo == TRUE),sum(python_remo == TRUE),
                  sum(csharp_remo==TRUE),sum(java_remo==TRUE))

df_remo <- data.frame("lenguajes" = lenguajes_remo,"valores"=valores_remo)
df_remo <-  cbind(df_remo,Modalidad="Remoto")

##HIBRIDO
#PHP
php_hibrid <- str_detect(desc_hibrido, regex("php|laravel", ignore_case = TRUE))

#Javascript
javascript_hibrid <- str_detect(desc_hibrido, regex("javascript|js|react", ignore_case = TRUE))

#python
python_hibrid <- str_detect(desc_hibrido, regex("python|django", ignore_case = TRUE))

#c#
csharp_hibrid <- str_detect(desc_hibrido, regex("c#", ignore_case = TRUE))


#java
java_hibrid <- str_detect(desc_hibrido,"(\\s+|,|^)[Jj]ava(\\s+|,|$)")



######### DATA FRAME #########
lenguajes_hibrid <- c("PHP","JavaScript","Python","C#","Java")
valores_hibrid <- c(sum(python_hibrid == TRUE),sum(javascript_hibrid == TRUE),sum(python_hibrid == TRUE
                  sum(csharp_hibrid==TRUE),sum(java_hibrid==TRUE))

df_hibrid <- data.frame("lenguajes"=lenguajes_hibrid,"valores"=valores_hibrid)
df_hibrid <-  cbind(df_hibrid,Modalidad="Hibrido")


###### MERGE DE LAS 3 DATA FRAMES ###################
df_final <- merge(x=df_pres,y=df_remo, by = "lenguajes")
```

```
df_temp <- rbind(df_pres, df_remo)
df_final <- rbind(df_temp,df_hibrid)
```

```
###GRAFICA####
ggplot(df_final, aes(x=lenguajes, y=valores,fill=Modalidad))+
  geom_bar(stat = "identity")+
  labs(title="Lenguajes más solicitados por Modalidad",
       x ="Lenguaje", y = "Ofertas de empleo")
```

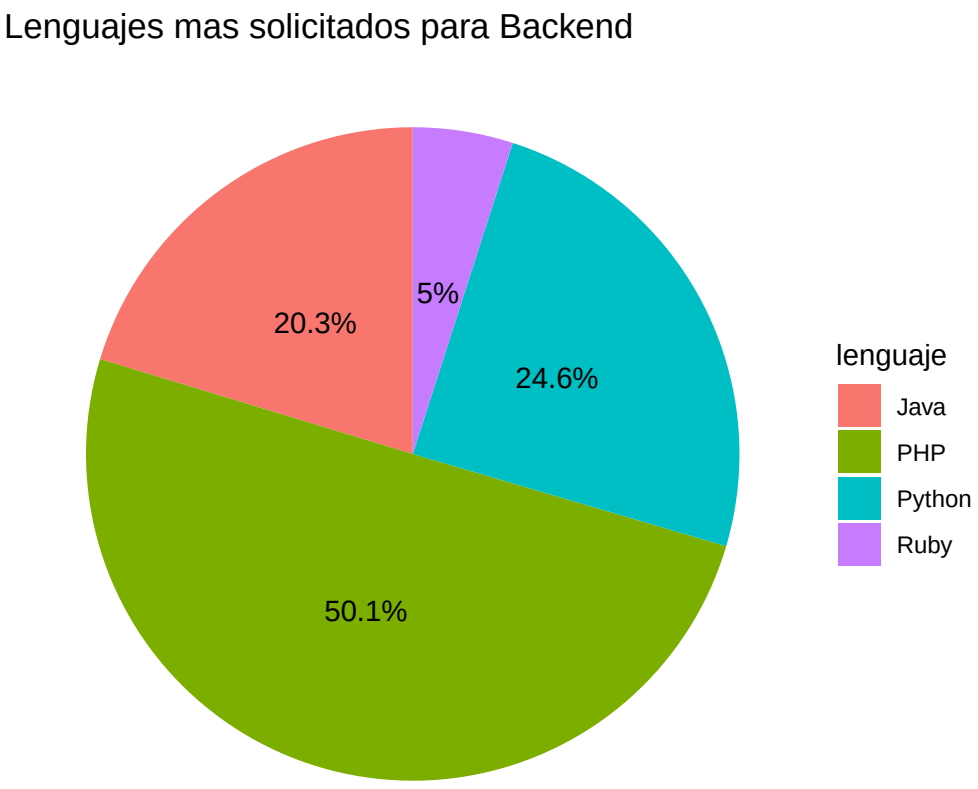


Pregunta 2 ¿Qué lenguajes de back-end son los más solicitados?

```
#Lenguajes backend
library(dplyr)
#PHP
php <- str_detect(descripcion, regex("php|laravel", ignore_case = TRUE))
phpCant<-sum(php == TRUE)
#Java
java <- str_detect(descripcion,"(\\s+|,|^)[Jj]ava(\\s+|,|$)")
javaCant<-sum(java==TRUE)
#Ruby
ruby <- str_detect(descripcion, regex("ruby", ignore_case = TRUE))
rubyCant<-sum(ruby == TRUE)
#Python
```

```r
python <- str_detect(descripcion, regex("python|django", ignore_case = TRUE))
pythonCant<-sum(python == TRUE)
```

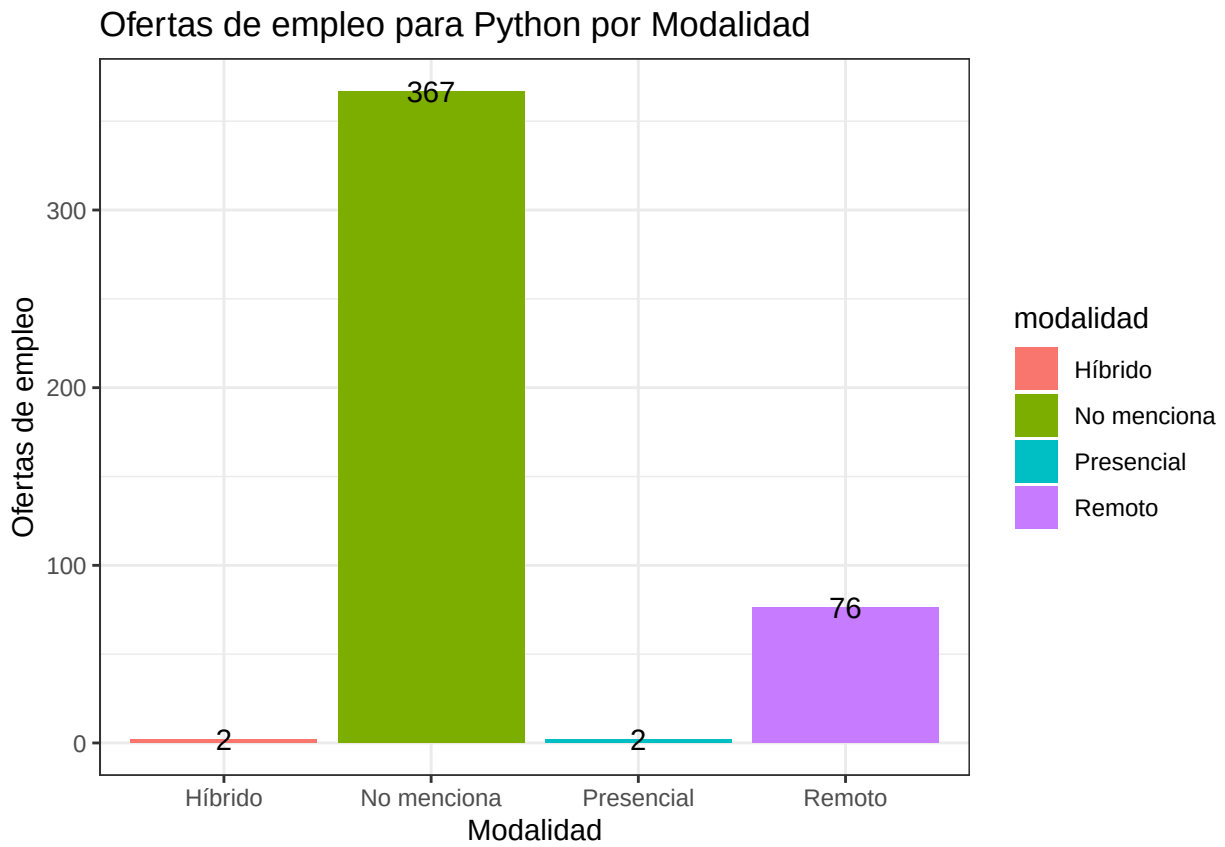
```r
valores_leng <-c(javaCant,phpCant,pythonCant,rubyCant)
datospr2 <- data.frame("lenguaje" = c("Java","PHP","Python","Ruby"),
                       "valor" = valores_leng)

#Porcentaje
porcentaje <- datospr2 %>%
  group_by(lenguaje) %>%
  count() %>%
  ungroup() %>%
  mutate(percentage=valores_leng/sum(valores_leng) * 100)

##Grafica de pastel ######

ggplot(porcentaje, aes(x="", y=percentage,fill=lenguaje))+
  geom_bar(stat="identity", width=1) +
  geom_text(aes(label = paste0(round(percentage,1),"%")),
            position = position_stack(vjust = 0.5)) +
  coord_polar(theta = "y") +
  labs(title= "Lenguajes mas solicitados para Backend")+
  theme_void()
```



Pregunta 3 ¿Para qué modalidad se solicita más el lenguaje Python?

```
python_presCant <- sum(python_pres==TRUE)
python_remoCant <- sum(python_remo==TRUE)
python_hibridCant <- sum(python_hibrid==TRUE)
python_mods <- python_presCant+python_remoCant+python_hibridCant

#no menciona modalidad##
python_noMenc <- pythonCant - python_mods

python_final <- data.frame("modalidad" = c("Presencial", "Remoto","Híbrido", "No menciona"),
                           "valor" = c(python_presCant,python_remoCant,python_hibridCant, python_noMenc
```

```
ggplot(python_final, aes(x=modalidad, y=valor, fill=modalidad)) +
  geom_bar(stat="identity")+
  geom_text(aes(label = valor))+
  theme_bw()+
  labs(title= "Ofertas de empleo para Python por Modalidad",
         x="Modalidad", y="Ofertas de empleo")
```

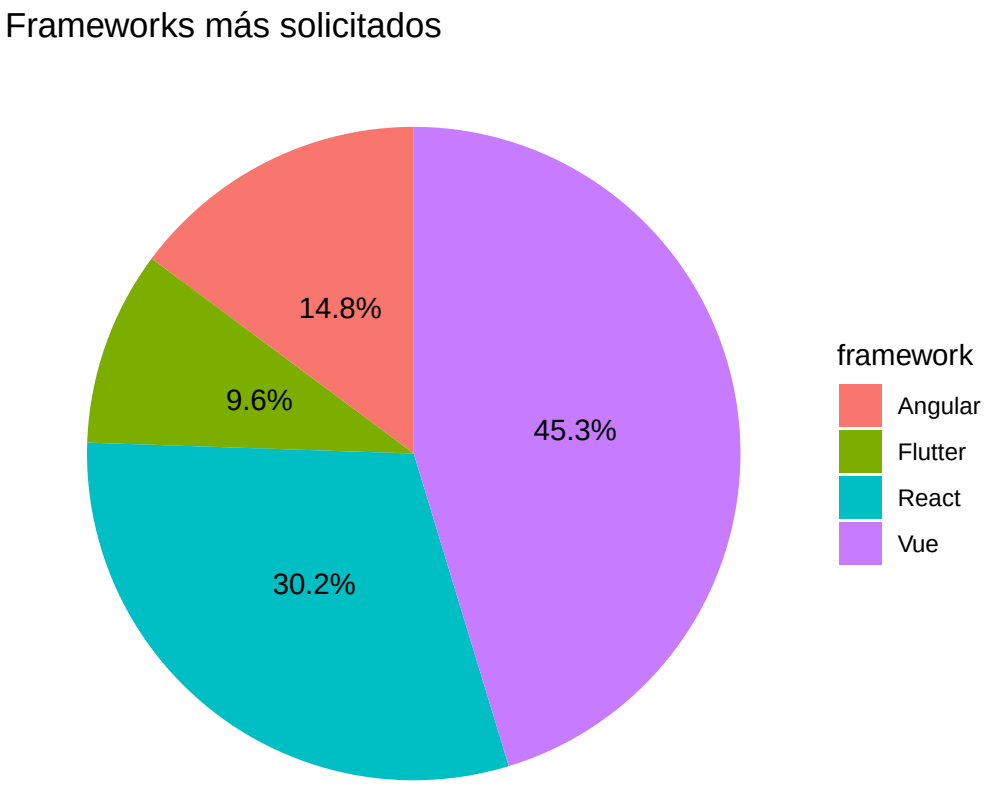


Pregunta 4 ¿Qué frameworks son los más solicitados?

```
#Lenguajes backend
library(dplyr)
#Vue
```

```
vue <- str_detect(descripcion, regex("vue|vue.js", ignore_case = TRUE))
vueCant<-sum(php == TRUE)
#Angular
angular <- str_detect(descripcion,regex("angular|angular.js", ignore_case= TRUE))
angularCant<-sum(angular==TRUE)
#Flutter
flutter <- str_detect(descripcion, regex("flutter", ignore_case = TRUE))
flutterCant<-sum(flutter == TRUE)

#React
react <- str_detect(descripcion, regex("react", ignore_case = TRUE))
reactCant<-sum(react == TRUE)
```

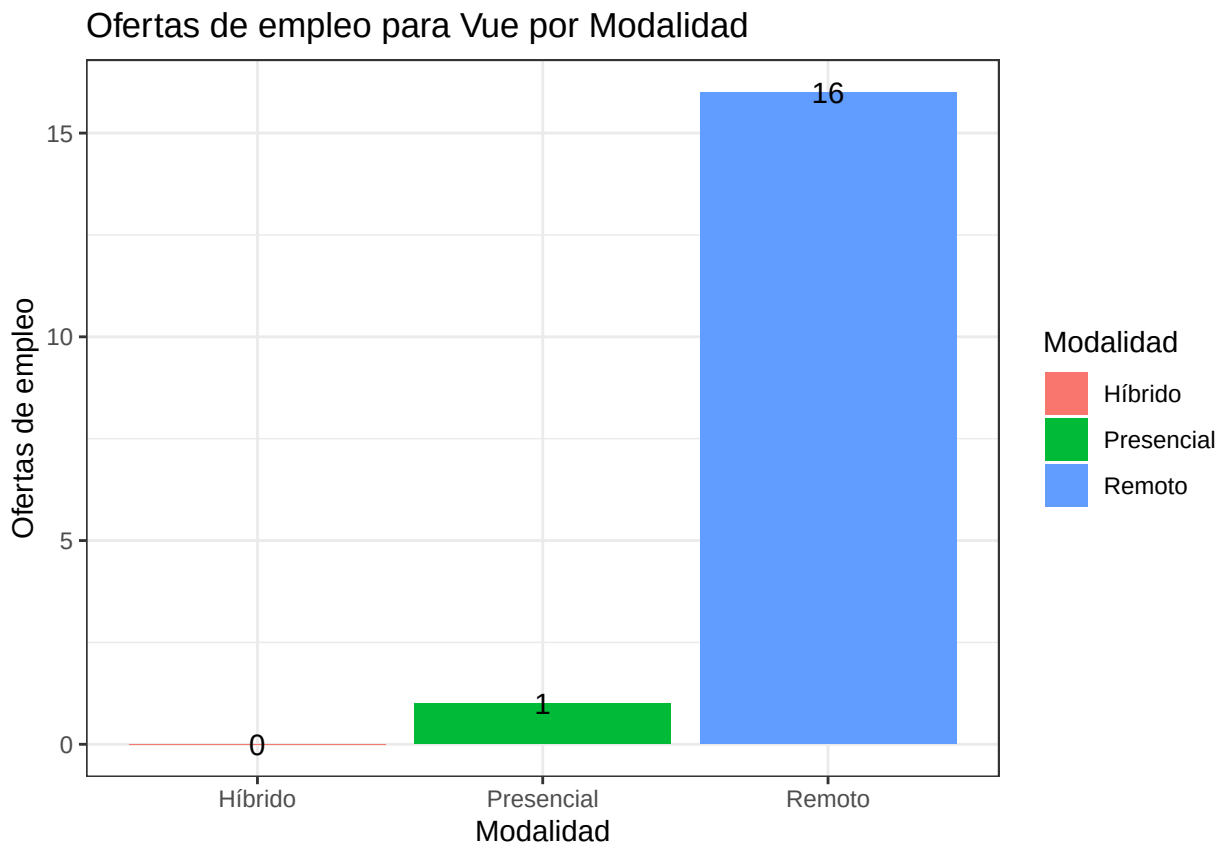
```
valores_frame <- c(angularCant,flutterCant,reactCant,vueCant)
datospr3 <- data.frame("framework" = c("Vue","Angular","Flutter","React"),
                       "valor" = valores_frame)
#Porcentaje
porcentaje_frame <- datospr3  %>%
  group_by(framework) %>%
  count() %>%
  ungroup() %>%
  mutate(percentage=valores_frame/sum(valores_frame) * 100)
##Grafica de pastel ######

ggplot(porcentaje_frame, aes(x="", y=percentage,fill=framework))+
  geom_bar(stat="identity", width=1) +
  geom_text(aes(label = paste0(round(percentage,1),"%")),
            position = position_stack(vjust = 0.5)) +
  coord_polar(theta = "y") +
  labs(title= "Frameworks más solicitados")+
  theme_void()
```



Pregunta 5 ¿En qué modalidad es más solicitado el framework Vue?

```
vue_pres <- str_detect(desc_presencial, regex("vue", ignore_case = TRUE))

vue_remo <-str_detect(desc_remota, regex("vue", ignore_case = TRUE))

vue_hibrid <-str_detect(desc_hibrido, regex("vue", ignore_case = TRUE))


vue_noMenc <- str_detect(descripcion, regex("vue", ignore_case = TRUE))



df_vue <- data.frame("Modalidad"= c("Presencial","Remoto","Híbrido"),
                     "Valores" = c(sum(vue_pres==TRUE),sum(vue_remo==TRUE),sum(vue_hibrid==TRUE)))
```

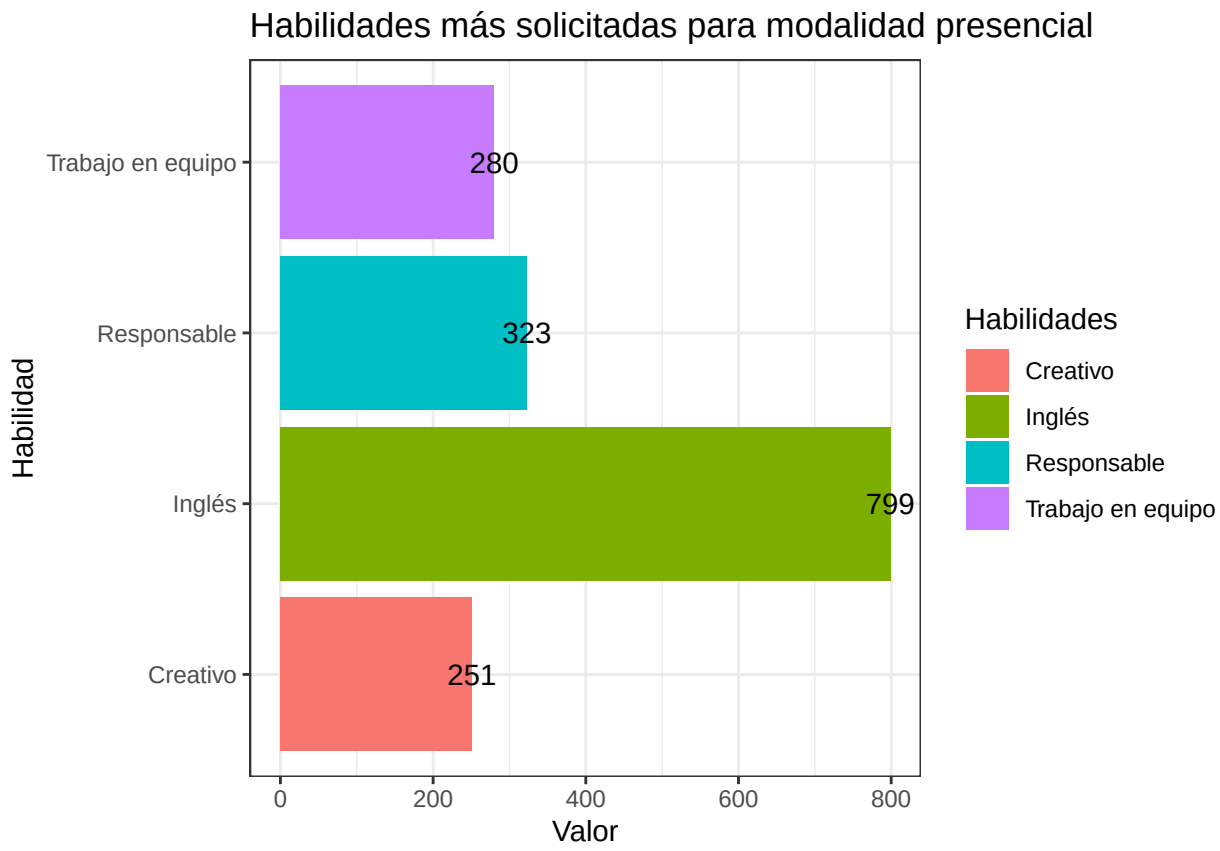
```
ggplot(df_vue, aes(x=Modalidad, y=Valores,fill=Modalidad)) +
  geom_bar(stat="identity")+
  geom_text(aes(label = Valores))+
  theme_bw()+
  labs(title= "Ofertas de empleo para Vue por Modalidad",
       x="Modalidad", y="Ofertas de empleo")
```



Pregunta 6 ¿Cuáles son las habilidades más solicitadas para la modalidad presencial?

```
responsabilidad <- str_detect(descripcion,regex("responsabilidad|responsability|responsable", ignore_cas
creativo <- str_detect(descripcion,regex("creatividad|creativo|creative|creativa", ignore_case = TRUE))
ingles <- str_detect(descripcion,regex("inglés|ingles|english", ignore_case = TRUE))
equipo <- str_detect(descripcion,regex("(\\s+|,|^)([Tt]eamwork|[Tt]rabajo en equipo|[Cc]olaboraci[oó]n)
df_habilidades <- data.frame("Habilidades" = c("Responsable", "Creativo","Inglés","Trabajo en equipo"),
                             "valor" = c(sum(responsabilidad == TRUE),sum(creativo == TRUE),sum(ingles =
                                         sum(equipo == TRUE)))
```

```
ggplot(df_habilidades, aes(x=Habilidades, y=valor,fill=Habilidades)) +
  geom_bar(stat="identity")+
  geom_text(aes(label = valor))+
  theme_bw()+
  labs(title= "Habilidades más solicitadas para modalidad presencial",
       x="Habilidad", y="Valor")+
   coord_flip()
```

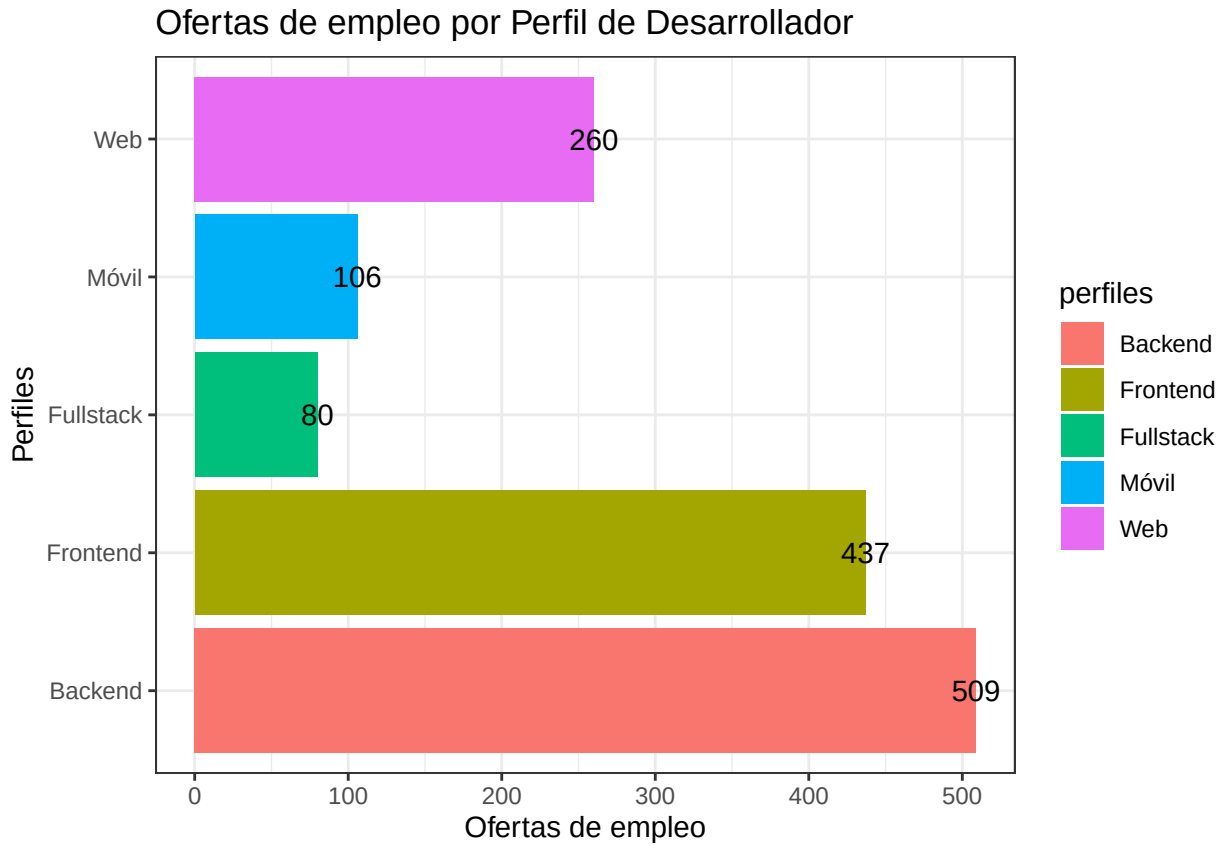


Pregunta 7 ¿Perfiles más buscados de desarrolladores?

```
#Desarrollador Frontend
frontend <- str_detect(descripcion,
regex("(\\s+|,|^)(([Dd]esarrollador(es)\\s*)?[Ff]ront-?end)|([Ff]ront-?end)\\s*([Dd]evelopers?))(\\s+|
ignore_case = TRUE))

#Desarrollador Backend
backend <- str_detect(descripcion,
regex("(\\s+|,|^)(([Dd]esarrollador(es)\\s*)?[Bb]ack-?end)|([Bb]ack-?end)\\s*([Dd]evelopers?))(\\s+|,|
ignore_case = TRUE))

#Desarrollador FullStack
fullstack <- str_detect(descripcion,
regex("(\\s+|,|^)(([Dd]esarrollador(es)\\s*)?[Ff]ullstack)|([Ff]ullstack)\\s*([Dd]evelopers?))(\\s+|,|
ignore_case = TRUE))

#Desarrollador web
web <- str_detect(descripcion,
regex("(\\s+|,|^)(([Dd]esarrollador(es))\\s*[Ww]eb)|([Ww]eb)\\s*([Dd]evelopers?))(\\s+|,|$)",
ignore_case = TRUE))

#Desarrollador movil
movil <- str_detect(descripcion,
```

```
regex("(\\s+|,|^)[Dd]esarroll(o|ador(es)?)\\s*[Dd]e\\s*[Aa]p(ps?|licaciones)\\s*([Mm][óo]vil(es)?)?(\\s
ignore_case = TRUE))
```

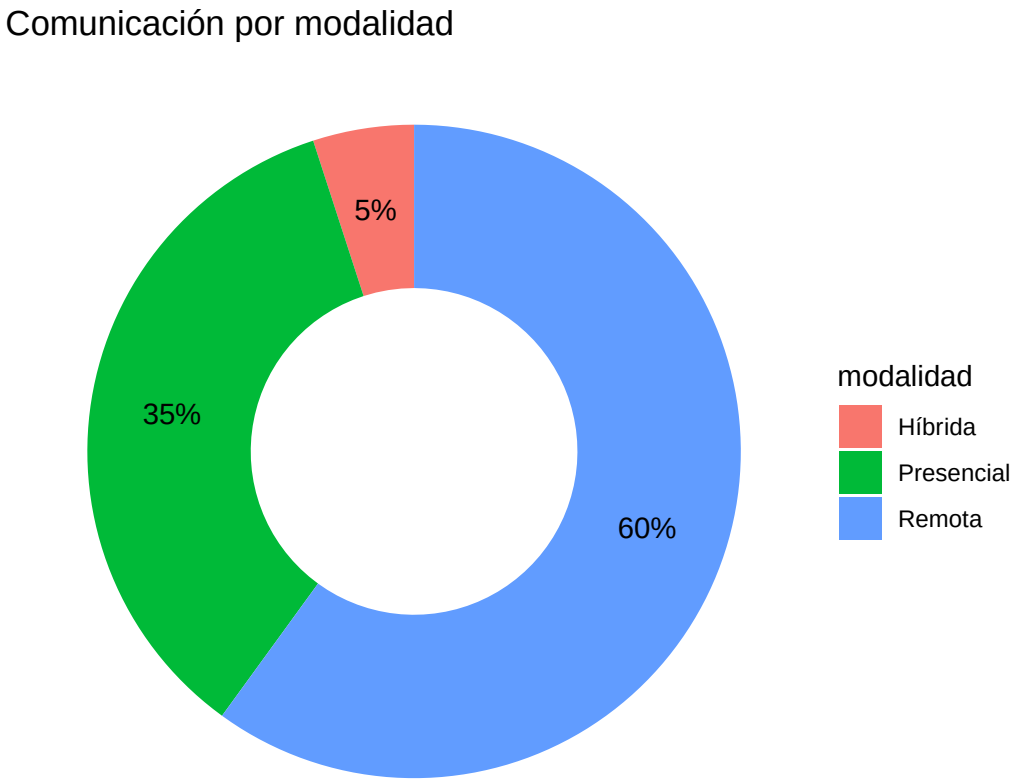
```
## Diagrama de barras ######
pregunta_7 <- data.frame("perfiles" = c("Frontend", "Backend","Fullstack", "Web", "Móvil"),
                         "valor" = c(sum(frontend==TRUE),sum(backend==TRUE),sum(fullstack==TRUE),

ggplot(pregunta_7, aes(x=valor, y=perfiles, fill=perfiles)) +
  geom_bar(stat="identity", orientation = "y")+
  geom_text(aes(label = valor))+
  theme_bw()+
  labs(title= "Ofertas de empleo por Perfil de Desarrollador",
       x="Ofertas de empleo", y="Perfiles")
```



Pregunta 8 ¿La habilidad "comunicación" es más requerida en modalidades remotas e híbrida que en modalidades presenciales?

```
com_pres <- str_detect(desc_presencial, regex("(\\s+|,|^)([Cc]omunicaci[oó]n|[Cc]omunication)(\\s+|,|$)

com_rem <- str_detect(desc_remota, regex("(\\s+|,|^)([Cc]omunicaci[oó]n|[Cc]omunication)(\\s+|,|$)", ig

com_hib <- str_detect(desc_hibrido, regex("(\\s+|,|^)([Cc]omunicaci[oó]n|[Cc]omunication)(\\s+|,|$)", ig
```

```r
## Gráfica de dona ######
valores<- c(sum(com_hib == TRUE),sum(com_pres == TRUE),sum(com_rem == TRUE))
datospr8 <- data.frame("modalidad" = c("Híbrida","Presencial","Remota"),
                       "valor" = valores)
porcentaje <- datospr8 %>%
  group_by(modalidad) %>%
  count() %>%
  ungroup() %>%
  mutate(percentage=valores/sum(valores) * 100)

ggplot(porcentaje,aes(x=2,y=percentage, fill=modalidad))+
  geom_bar(stat = "identity", width=1)+
  geom_text(aes(label = paste0(round(percentage,1),"%")),
            position = position_stack(vjust = 0.5))+
  coord_polar(theta = "y")+
  labs(title="Comunicación por modalidad")+
  theme_void()+
  xlim(0.5,2.5)
```



Pregunta 9 ¿Saber inglés es importante para desenvolverse en el ámbito del desarrollo de software?

```r
## Menciones de inglés
ingles <- str_detect(descripcion, regex("(\\s+|,|^)([Ii]ngl[eé]s|[Ee]nglish)(\\s+|,|$)", ignore_case = T
ingles <- sum(ingles == TRUE)
no_ingles = (length(descripcion) - sum(ingles == TRUE))
#no_ingles

## Gráfica de pastel ######
valores<- c(ingles, no_ingles)
datospr9 <- data.frame("label" = c("Inglés","No inglés"),
                       "valor" = valores)
porcentaje <- datospr9 %>%
  group_by(label) %>%
  count() %>%
  ungroup() %>%
  mutate(percentage=valores/sum(valores) * 100)

ggplot(porcentaje, aes(x="", y=percentage, fill=label))+
  geom_bar(stat="identity", width=1) +
  geom_text(aes(label = paste0(round(percentage,1),"%")),
            position = position_stack(vjust = 0.5)) +
  coord_polar(theta = "y") +
  labs(title= "Inglés en el ámbito del desarrollo de software")+
  theme_void()
```