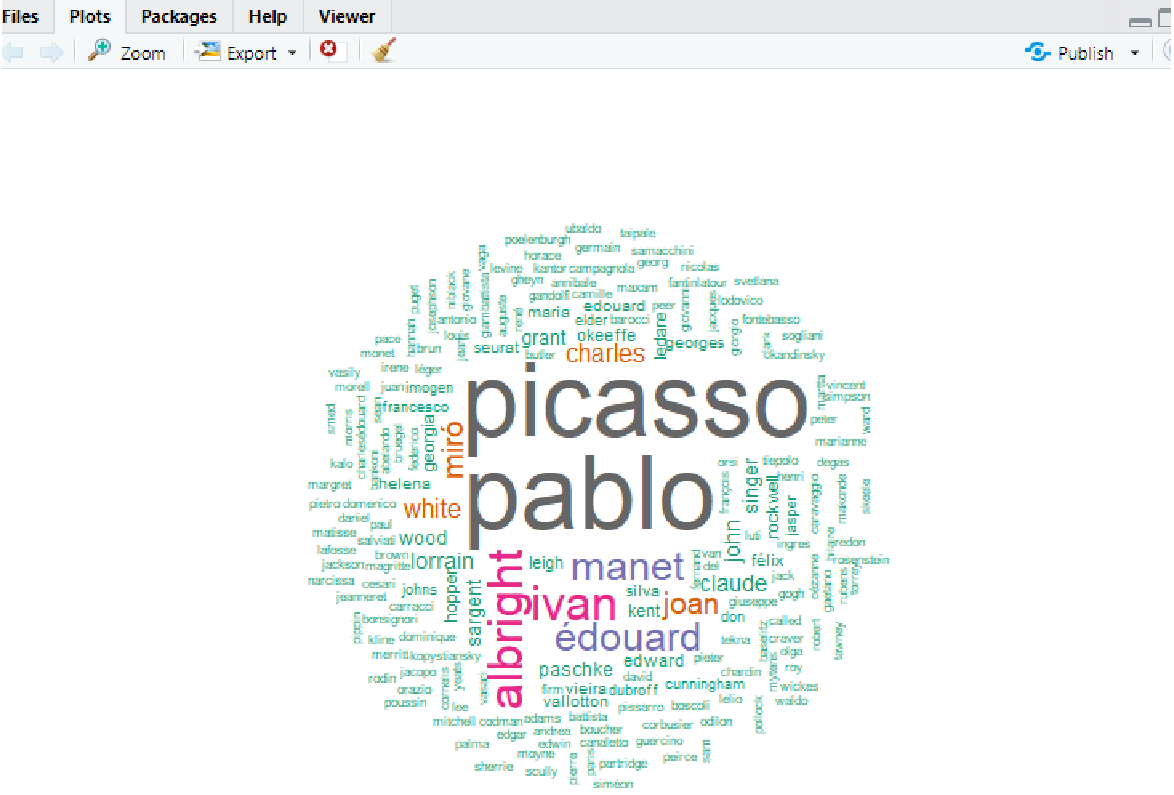
During the processing of the data, we had difficulty in extracting it, since everything was saved as a list and not as a dataframe. This did not allow us to work with ggplot, even when trying to reconvert the generated list of strings, for example using the command "(as.data.frame(data))". Moreover, even using the command "(typeoff(data))", the response received was still the same. the answer received was still "lists". We therefore decided to turn our analysis towards the creation of several wordclouds.We have decided to create 3 wordclouds, to do this we have always used the same procedure, changing only the variable through which to conduct the analysis. It has always been a matter of extracting the inherent column, first create the corpus, clean it of punctuation and other signs not appreciated within the analysis, create the following Document Term Matrix, and from there through the wordcloud library, create our word cloud. Below is an explanation of each step after the # symbol.

```
install.packages("tm")

install.packages("tidyverse")

install.packages("wordcloud")

install.packages("readr")


#we re_call the libraries that we need


library(readr)

library(tm)

library(wordcloud)

library(tidyverse)


#importing the csv file

data <- read_csv("art1000.csv")



#we decided to create 3 wordcloud (cloud with words, the more the are present, the more they'll appear in the cloud)

#first cloud, the most present Artist

#name's extraction from the dataset

artist_title <- (data$artist_title)



#create the corpus
```

```
docs <- VCorpus(VectorSource(artist_title))



#clean the corpus
clean_corpus <- function(CORPUS){
  CORPUS <- tm_map(CORPUS, content_transformer(tolower))
  CORPUS <- tm_map(CORPUS, content_transformer(removeWords), stopwords("en"))
  CORPUS <- tm_map(CORPUS, content_transformer(removePunctuation))
  CORPUS <- tm_map(CORPUS, content_transformer(removeNumbers))
  CORPUS <- tm_map(CORPUS, content_transformer(stripWhitespace))
  return(CORPUS)
}

correct_artist <- clean_corpus(docs)

#creation of the dtm
dtm <- TermDocumentMatrix(correct_artist)
matrix <-(as.matrix(dtm))
words <- sort(rowSums(matrix),decreasing=TRUE)
df <- data.frame(word = names(words),freq=words)

#performing the first wordcloud
set.seed(1234)
wordcloud(words = df$word, freq = df$freq, min.freq = 1,          max.words=200, random.order=FALSE,
rot.per=0.35,          colors=brewer.pal(8, "Dark2"))
```



#Pablo Picasso is the one with the most attendances

#now we want to know, the most used title's word

#we extract the title column

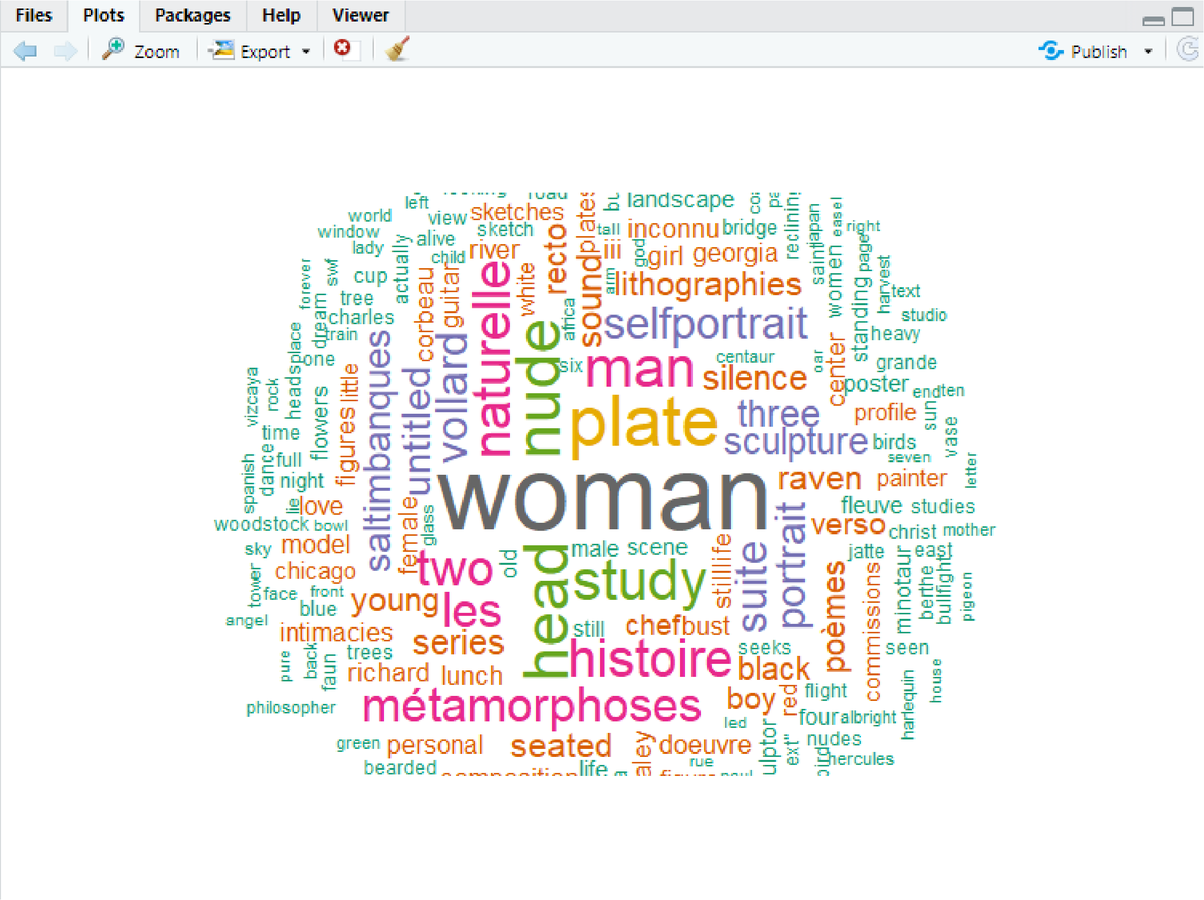
title <- (data$title)

#create the corpus

docs1<- VCorpus(VectorSource(title))

#clean the corpus

```
clean_corpus <- function(CORPUS){
  CORPUS <- tm_map(CORPUS, content_transformer(tolower))
  CORPUS <- tm_map(CORPUS, content_transformer(removeWords), stopwords("en"))
  CORPUS <- tm_map(CORPUS, content_transformer(removePunctuation))

  CORPUS <- tm_map(CORPUS, content_transformer(removeNumbers))

  CORPUS <- tm_map(CORPUS, content_transformer(stripWhitespace))

  return(CORPUS)

}


correct_title <- clean_corpus(docs1)


#creation of the dtm

dtm2<- TermDocumentMatrix(correct_title)

matrix2 <-(as.matrix(dtm2))

words2 <- sort(rowSums(matrix2),decreasing=TRUE)

df2 <- data.frame(word = names(words2),freq=words2)


#performing the second wordcloud

set.seed(1234)

wordcloud(words = df2$word, freq = df2$freq, min.freq = 1,          max.words=200, random.order=FALSE,
rot.per=0.35,          colors=brewer.pal(8, "Dark2"))
```



#woman is the most frequent title's word, we could expext a lot of works, representing woman

#as the third wordcloud we want to know, what is the nation the gives the museum the majority of the works

#we extract the relative column from the dataset

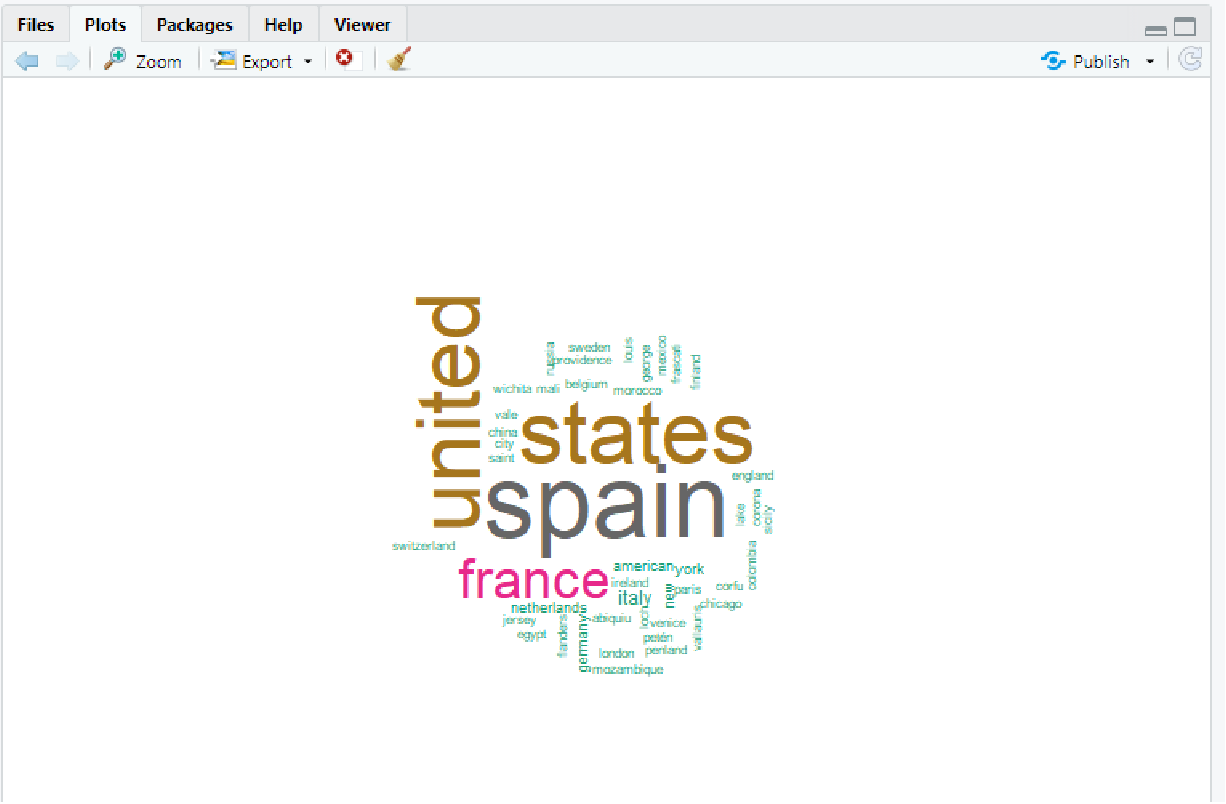
```
piace_of_origins <- (data$place_of_origin)
```

#create the corpus

```
docs3 <- VCorpus(VectorSource(piace_of_origins))
```

#clean the corpus

```
clean_corpus <- function(CORPUS){
  CORPUS <- tm_map(CORPUS, content_transformer(tolower))
  CORPUS <- tm_map(CORPUS, content_transformer(removeWords), stopwords("en"))
  CORPUS <- tm_map(CORPUS, content_transformer(removePunctuation))
  CORPUS <- tm_map(CORPUS, content_transformer(removeNumbers))
  CORPUS <- tm_map(CORPUS, content_transformer(stripWhitespace))
  return(CORPUS)
}

correct_origin <- clean_corpus(docs3)

#the new dtm
dtm3<- TermDocumentMatrix(correct_origin)
matrix3 <-(as.matrix(dtm3))
words3 <- sort(rowSums(matrix3),decreasing=TRUE)
df3 <- data.frame(word = names(words3),freq=words3)

#we have our third cloud
set.seed(1234)
wordcloud(words = df3$word, freq = df3$freq, min.freq = 1,          max.words=1000, random.order=FALSE,
rot.per=0.35,           colors=brewer.pal(8, "Dark2"))
```



#the most of the painting comes from United States and Spain