Assignment 4

Inshal Naqvi

##Question 1

This problem will involve the nycflights13dataset(including tables flights, airlines,airports,planes andweather), which we saw in class. It is available in both R and Python, however R is recommended for at least the visualization portion of the question. You can get more information about this package on github at https://github.com/tidyverse/nycflights13 The data tables can be found in the data-raw folder of the above-mentioned github repository. Additionally, the flights.csv file which was used in assignment 3 is available in the Datasets module on Canvas. Start by installing and importing the dataset to your chosen platform. We will first use joins to search and manipulate the dataset, then we will produce a flightcountvisualization.

```
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

library('tidyverse')

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.
3.1 ──

## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.1.4     ✓ stringr 1.4.0
## ✓ tidyr   1.1.3     ✓ forcats 0.5.1
## ✓ readr   2.0.2

## ── Conflicts ────────────────────────────────────────── tidyverse_conflict
s() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(nycflights13)
library(tidyverse)

data("airlines")
```

```r
data("airports")
data("flights")
data("planes")
data("weather")
```

##Part a

Filter the dataset(using a left join) to display the tail number, year, month, day, hour, origin, and humidity for all flights heading to Tampa International Airport(TPA) after 12p mon November 1, 2013.

```r
flights_to_tpa = left_join(flights, weather) %>%
  select(tailnum, year, month, day, hour, origin, humid, dest)%>%
  filter(year == 2013 & month == 11 & day == 1 & hour >= 12 & dest == "TPA")

## Joining, by = c("year", "month", "day", "origin", "hour", "time_hour")

flights_to_tpa

## # A tibble: 10 × 8
##    tailnum  year month   day  hour origin humid dest 
##    <chr>   <int> <int> <int> <dbl> <chr>  <dbl> <chr>
##  1 N580JB   2013    11     1    14 JFK     63.1 TPA  
##  2 N337NB   2013    11     1    14 LGA     56.5 TPA  
##  3 N567UA   2013    11     1    15 EWR     52.8 TPA  
##  4 N515MQ   2013    11     1    14 JFK     63.1 TPA  
##  5 N779JB   2013    11     1    15 EWR     52.8 TPA  
##  6 N561JB   2013    11     1    16 LGA     50.6 TPA  
##  7 N974DL   2013    11     1    18 JFK     74.8 TPA  
##  8 N319NB   2013    11     1    19 LGA     60.5 TPA  
##  9 N76265   2013    11     1    19 EWR     72.5 TPA  
## 10 N768JB   2013    11     1    19 JFK     83.5 TPA
```

#Part b

What is the difference between the following two joins?

```r
anti_join_1 = anti_join(flights, airports, by = c("dest" = "faa"))
anti_join_2 = anti_join(airports, flights, by = c("faa" = "dest"))
```

According to the scenario of nycflights13, the first Anti_Join will show all those flights that have a destination to to those Airports which are not listed in the original Airports list where flights$dest = airports$faa.

The second Anti_join in which the primary dataset is airports will show those airports and there names which are either either not operational and flights does not operate or there were no flights to those airports in 2013.

#Part c

Filter the table flights to only show flights with planes that have flown at least 100 flights. Hint: tailnum is used to identify planes.(suggested functions: R:semi_join(), count(), filter();

```
#Filtering missing tail number
planes_gte100 <- flights %>%
  filter(!is.na(tailnum)) %>%
  group_by(tailnum) %>%
  count() %>%
  filter(n >= 100)

#Semi-join planes that have flown at 100 flights

flights %>%
  semi_join(planes_gte100, by = "tailnum")
```

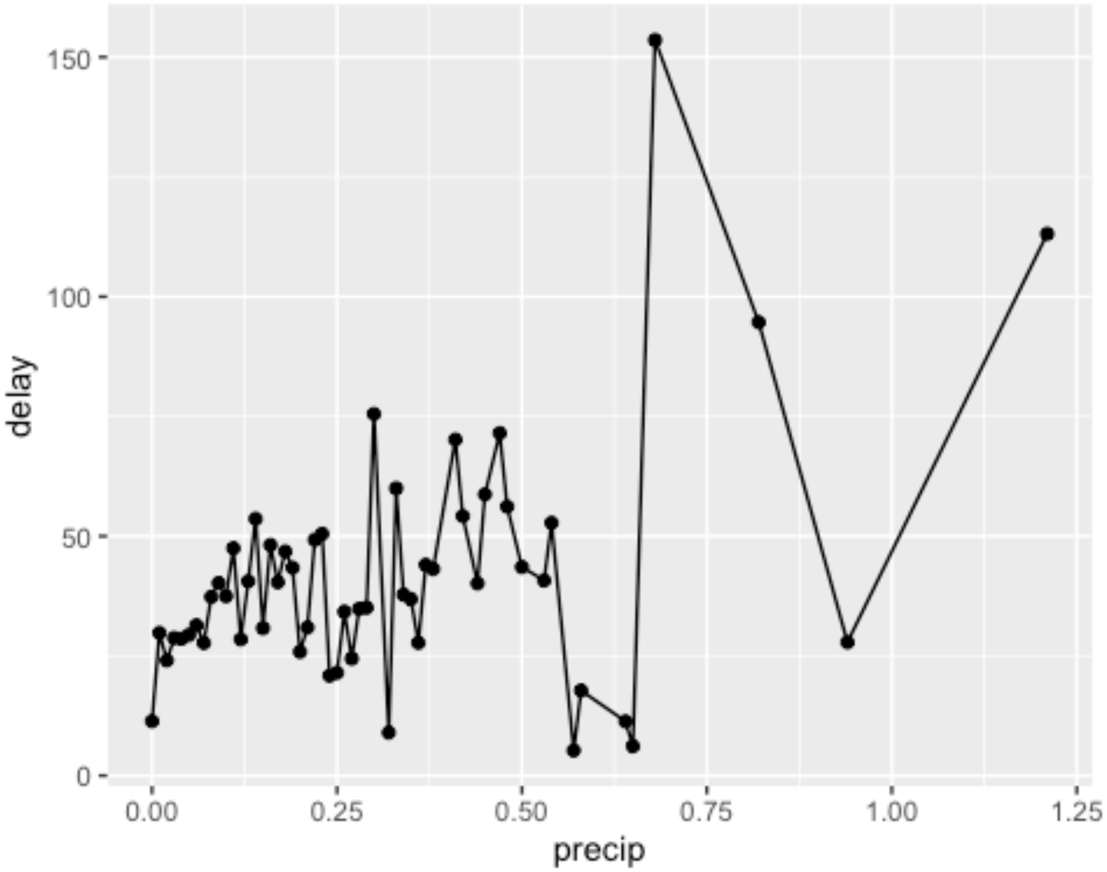
```
## # A tibble: 228,390 × 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_
time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <
int>
##  1  2013     1     1      517            515         2      830
819
##  2  2013     1     1      533            529         4      850
830
##  3  2013     1     1      544            545        -1     1004
1022
##  4  2013     1     1      554            558        -4      740
728
##  5  2013     1     1      555            600        -5      913
854
##  6  2013     1     1      557            600        -3      709
723
##  7  2013     1     1      557            600        -3      838
846
##  8  2013     1     1      558            600        -2      849
851
##  9  2013     1     1      558            600        -2      853
856
## 10  2013     1     1      558            600        -2      923
937
## # … with 228,380 more rows, and 11 more variables: arr_delay <dbl>,
## #   carrier <chr>, flight <int>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time_hour <d
ttm>
```

#Part d

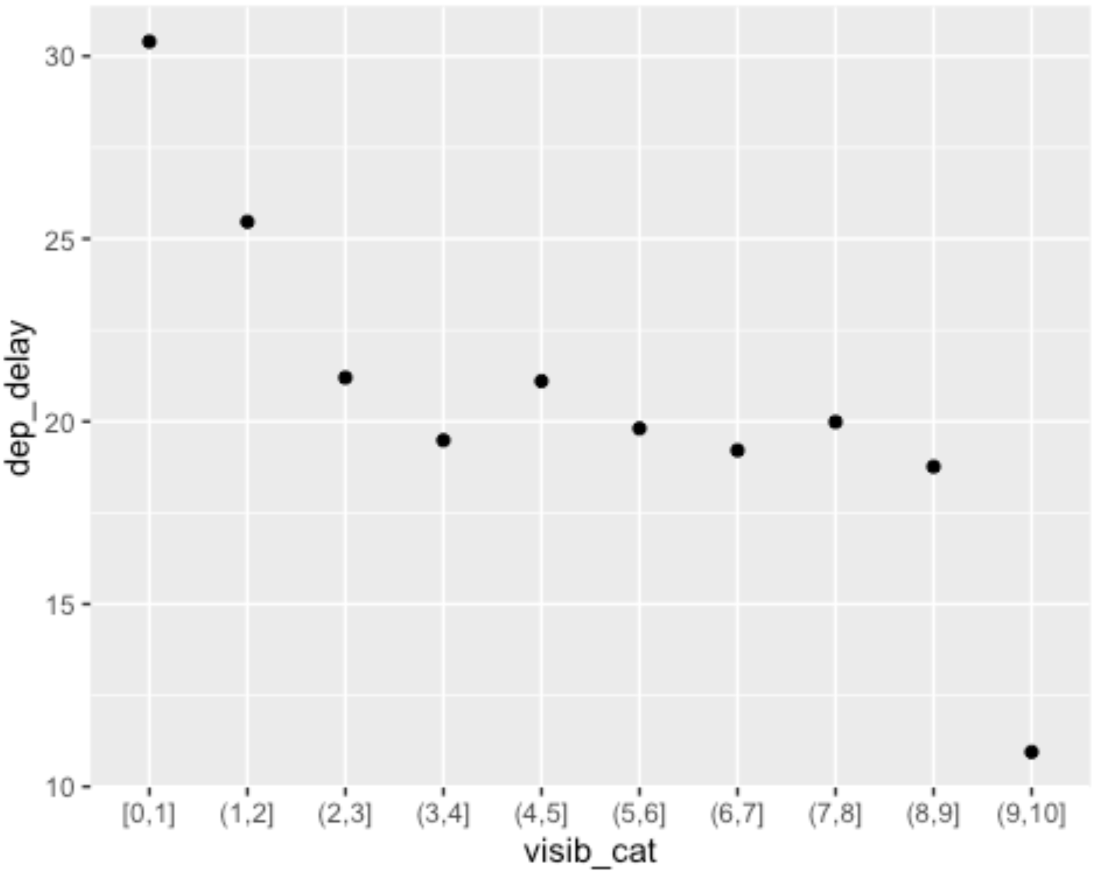
What weather conditions make it more likely to see a delay? Briefly discuss any relations/patterns you found.

```
flight_weather <-
  flights %>%
  inner_join(weather, by = c(
```

```r
    "origin" = "origin",
    "year" = "year",
    "month" = "month",
    "day" = "day",
    "hour" = "hour"
  ))

flight_weather %>%
  group_by(precip) %>%
  summarise(delay = mean(dep_delay, na.rm = TRUE)) %>%
  ggplot(aes(x = precip, y = delay)) +
  geom_line() + geom_point()
```



```r
flight_weather %>%
  ungroup() %>%
  mutate(visib_cat = cut_interval(visib, n = 10)) %>%
  group_by(visib_cat) %>%
  summarise(dep_delay = mean(dep_delay, na.rm = TRUE)) %>%
  ggplot(aes(x = visib_cat, y = dep_delay)) +
  geom_point()
```



We see a delay whenever there is precepatation unless its above 0.02 then the trend is not strong.

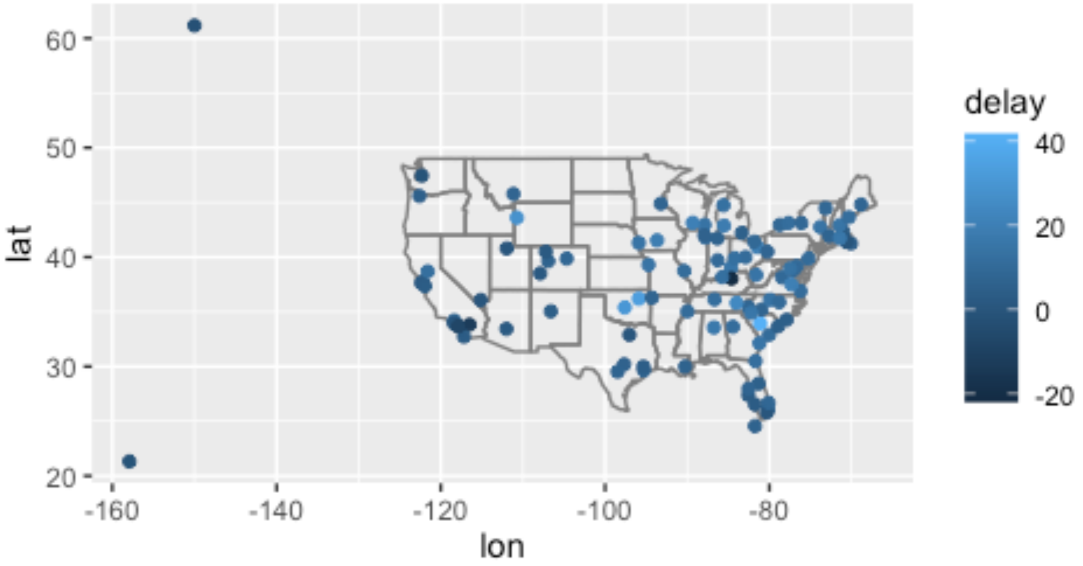
There is a strong relationship between delays and visibility. If visibility is less than 2 miles delays are higher when also agrees intuitively.

#Part e

Produce a map that sizes each destination airport by the number of incoming flights.You may use a continuous scale for the size. Here is a code snippet to draw a map of all flight destinations, which you can use as a starting point. You may need to install the maps packages if you have not already. Adjust the title, axis labels and aesthetics to make this visualization as clear as possible.

```
library(tidyverse)
library(sf)

## Linking to GEOS 3.8.1, GDAL 3.2.1, PROJ 7.2.1

library(here)

## here() starts at /Users/enshalnaqvi/Desktop/cpts575/Assignment_4

library(ggplot2)
```

```
airports_1e = flights%>%
  group_by(dest) %>%
  # arrival delay NA's are cancelled flights
  summarise(delay = mean(arr_delay, na.rm = TRUE)) %>%
  inner_join(airports, by = c(dest = "faa"))

airports_1e%>%
  ggplot(aes(lon, lat, colour = delay)) +
  borders("state") +
  geom_point() +
  coord_quickmap()
```



#Problem 2 Tableau screenshot are submitted.

```
covid19 = read.csv("covid19_vaccinations_USA.csv", sep=",", header = TRUE)
#head(covid19)

janssen = covid19[c(1:2)]
#head(janssen)

moderna = covid19[c(1,3)]
#head(moderna)
```

```
pfizer = covid19[c(1,4)]
#head(pfizer)
```

Question3

```
library("tm")

## Loading required package: NLP

##
## Attaching package: 'NLP'

## The following object is masked from 'package:ggplot2':
##
##     annotate

library("SnowballC")
library("wordcloud")

## Loading required package: RColorBrewer

library("RColorBrewer")

text_q3 = read.delim("getting_started_with_ml.txt")
data_q3 = Corpus(VectorSource(text_q3))

word_cloud = content_transformer(function (x , pattern ) gsub(pattern, " ", x
))

data_q3 = tm_map(data_q3, word_cloud, "/")

## Warning in tm_map.SimpleCorpus(data_q3, word_cloud, "/"): transformation d
rops
## documents

data_q3 = tm_map(data_q3, word_cloud, "@")

## Warning in tm_map.SimpleCorpus(data_q3, word_cloud, "@"): transformation d
rops
## documents

data_q3 = tm_map(data_q3, word_cloud, "\\|")

## Warning in tm_map.SimpleCorpus(data_q3, word_cloud, "\\|"): transformation
drops
## documents

data_q3 = tm_map(data_q3, word_cloud, "the")

## Warning in tm_map.SimpleCorpus(data_q3, word_cloud, "the"): transformation
drops
## documents

data_q3 = tm_map(data_q3, word_cloud, "you")
```

```
## Warning in tm_map.SimpleCorpus(data_q3, word_cloud, "you"): transformation
drops
## documents
```

```
text_mine = TermDocumentMatrix(data_q3)
m = as.matrix(text_mine)
v = sort(rowSums(m),decreasing=TRUE)
d = data.frame(word = names(v),freq=v)
head(d, 10)
```

```
##                  word freq
## learning     learning   32
## are               are   24
## data             data   23
## machine       machine   23
## that             that   23
## for               for   21
## can               can   18
## "•                 "•   16
## learning", learning",   16
## will             will   16
```

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

```
## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : "step could not be fit on page. It will not be plotted.
```

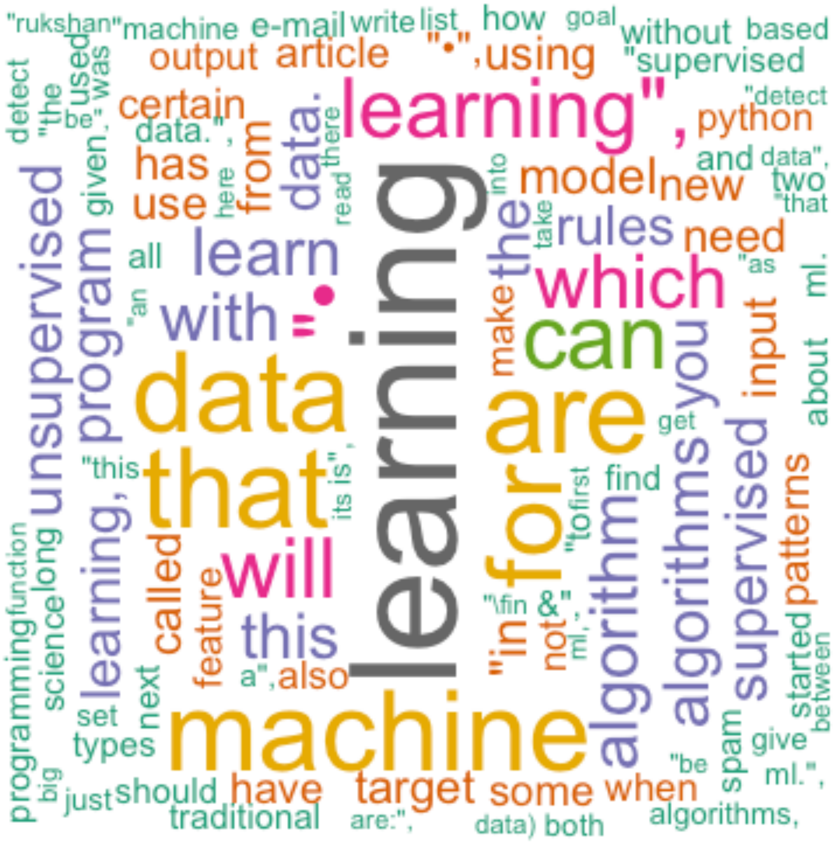
```
## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : (called could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : can", could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : changes could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : complex could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : following could not be fit on page. It will not be plotted.
## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : getting could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : label could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : many could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : mapping could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : process. could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : supervise could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : techniques could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : "algorithm could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : "reinforcement could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : "there could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : "unsupervised could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : (ml)", could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : adapt could not be fit on page. It will not be plotted.
## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : although could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : article", could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : automatically could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : becomes could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : categories could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : classical could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : classification could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : classified could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : code could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : compared could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : consists could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : dataset could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : disruptive could not be fit on page. It will not be plotted.
## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : done could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : don’t could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : effectively could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : else, could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : emails could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : february could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : filter could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : given. could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : help could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : human could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : implement could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : learned. could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : learning. could not be fit on page. It will not be plotted.
## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : learning?", could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : main could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : meet could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : method could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : necessary could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : points could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : pramoditha", could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : predict could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : prediction could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : process could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : produce could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : regression could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : rukshan could not be fit on page. It will not be plotted.
## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : rules. could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : selected could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : semi-supervised could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : simple could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : single could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : software could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : solution could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : solutions could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : solve could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : soon could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : statistics could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : statistics, could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : successfully could not be fit on page. It will not be plotted.
## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : table could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : terms could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : the", could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : training could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : type could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : under could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : useful could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : value could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : variable could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : variables could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : very could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : what could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : when", could not be fit on page. It will not be plotted.
## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : where could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : work could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : would could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.word
s =
## 200, : “for could not be fit on page. It will not be plotted.
```



```
#title(main = "Word Cloud - Introduction to ML", font.main = 1, cex.main = 1)
```