Marwa El Awik 2022-11-15 SGSF Executive Director Name: Lorettta Williams Gurnell SGSF Executive Director Signature:

Date: 6 Dec 2022

Code

Code **▼**

About Dataset

The dataset refers to 18 SUPERGirls of SGSF Foundation. It includes data collected from a survey filled out by those Girls. Every SUPERGirl provided us with her first Name, last Name, home Address, email, cell phone number, date of birth, race or ethnicity, age, current grade, the name of the school that she is currently attending, her preferred mentor, her professional interest or expertise, if she has or has not reliable transportation in case she needs to meet her mentor, if she had a mentoring role before, her language proficiency, her classification, her major/minor or certifications, also, she answered many questions.

Looking ahead, of interest for the SUPERGirls SHINE Foundation will be to apply a sentiment analysis based on the SUPERGirls expectations from the mentorship program.

Here is a glimpse of what we will be working with.

```
Code
## # A tibble: 6 x 29
      `Submission Date` `First Name` `Last Name` `Home Address`
                                                                                   `The name of t^{\sim}
##
     <chr>
                           <chr> <chr>
                                                           <chr>
                                                                                   <chr>
## 1 27-Oct-22 Yaquelin Rodriguez <NA> University of H-
## 2 23-Oct-22 Jennifer Li <NA> Rice University
## 3 18-Oct-22 Joauna Carter 18003 Oak Cottage~ Harmony School ~
## 4 17-Oct-22 Nadiya Kabugu 13130 Fry Rd. Cyp~ Smith Middle Sc~
## 5 13-Oct-22 breida ulibarri <NA> <NA>
                    Daniella
                                            Garcia
                                                          7214 Calais Rd. H~ Kipp Sunnyside
## 6 13-0ct-22
## # ... with 24 more variables: `Date of Birth` <chr>, `Race or Ethnicity` <chr>,
## # Age <dbl>, `Your Current Grade` <chr>,
        `After school, do you have reliable transportation, if you need to meet your mentor?` <chr>,
## #
         `Language Proficiencies. List all.` <chr>,
## #
## #
         `Your Hobbies and Interests` <chr>, `What Best Describes you` <chr>,
## #
         `Your classification.` <chr>,
         `If in college, your major/minor and any certifications.` <chr>, ...
## #
```

For this project, we are no going to dive deeper into cleaning the data, our main focus will be on the text columns only and particularly on the question: "What are you expectations in this mentoring program"

In the first step, we will extract this text column from the dataset. Then we will start cleaning it by removing the undesirable words, especially mentioned in the question: "mentor", "Mentor", "Mentoring", "Program", "program".

After that, we will extract the words out of the sentences, and get only the words other then the undesired words, also we will get rid of the stop words using the "tidytext package" and those words that are less than 3 words such as "I", and finally we will get the counts of the words

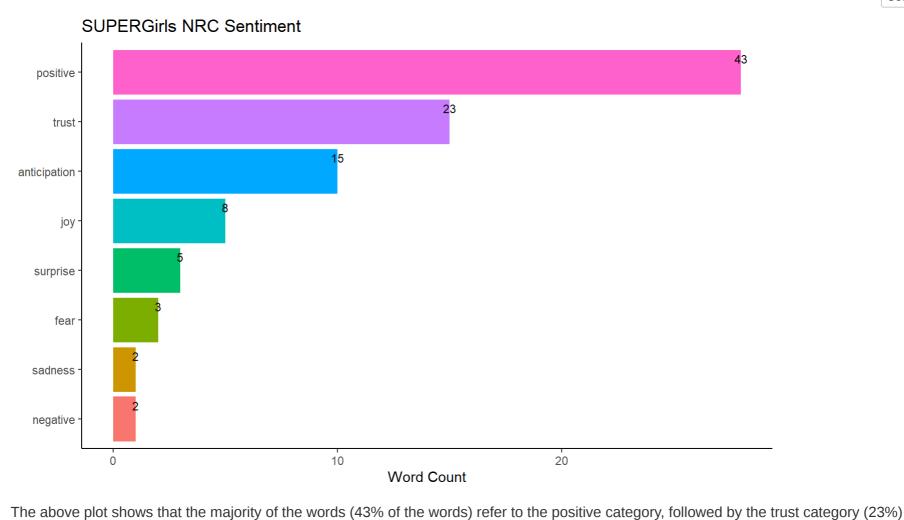
```
Code
## # A tibble: 10 x 2
##
     word
     <chr>
                  <int>
## 1 expect
## 2 field
                       5
## 3 learn
   4 career
## 5 engineering
   6 expand
                       3
   7 experience
  8 opportunities
                       3
  9 study
                       3
## 10 choose
```

Now, we will assign every word into NRC lexicon categories: positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise, and trust by inner joining the words that we got previously with the NRC sentiments dataset provided by the tidytext package.

Code

```
## # A tibble: 65 x 3
##
     word
                     n sentiment
##
     <chr>
                  <int> <chr>
                   5 anticipation
## 1 expect
   2 expect
                     5 positive
                     5 surprise
   3 expect
   4 expect
                   5 trust
                   5 positive
   5 learn
   6 career
                     4 anticipation
   7 career
                     4 positive
                      3 positive
  8 study
  9 knowledge
                     2 positive
## 10 professional
                     2 positive
## # ... with 55 more rows
```

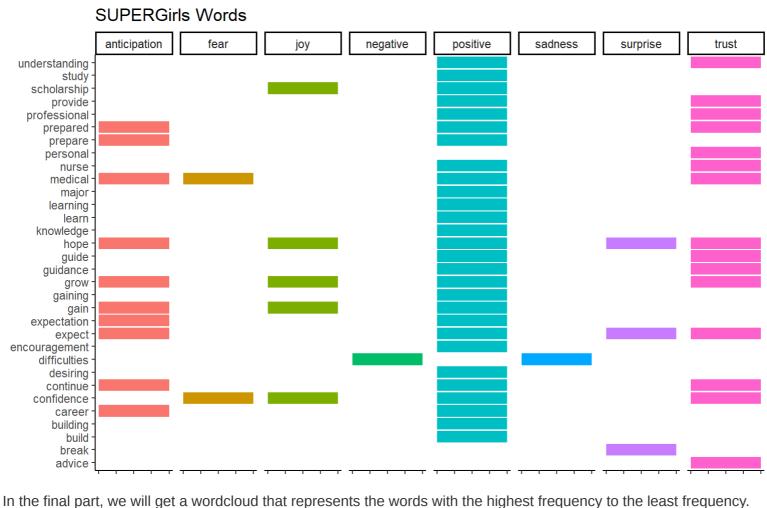
Lets visualize our SUPERGirls sentiments.



and the anticipation (15%). Only 2% of the words used by the SGSF girls refer to sadness and 2% refers to negative feelings.

Lets look at the words for each category.

Code



n the linal part, we will get a wordcloud that represents the words with the highest frequency to the least frequenc

pediatricprovide planned industry growing personalexposure types internship expectation confidence matter break i'm fit add study difficulties quide advice tech biology build build desiring book interpolation learning knowledge exploring specific potentially encouragement life major support insight listener scholarship teaches prepare

Code

Code