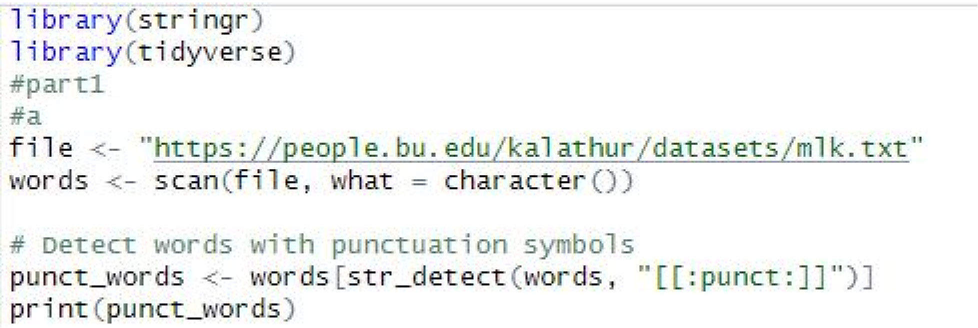
Weilin Lu

METCS566

Assignment 6

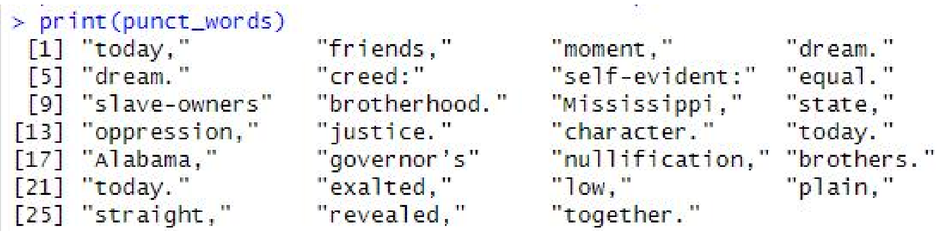
Part1

A) :



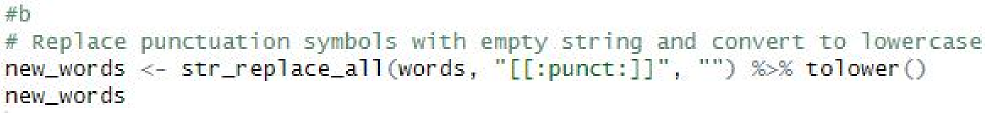
```
library(stringr)
library(tidyverse)
#part1
#a
file <- "https://people.bu.edu/kalathur/datasets/mlk.txt"
words <- scan(file, what = character())

# Detect words with punctuation symbols
punct_words <- words[str_detect(words, "[[:punct:]]")]
print(punct_words)
```

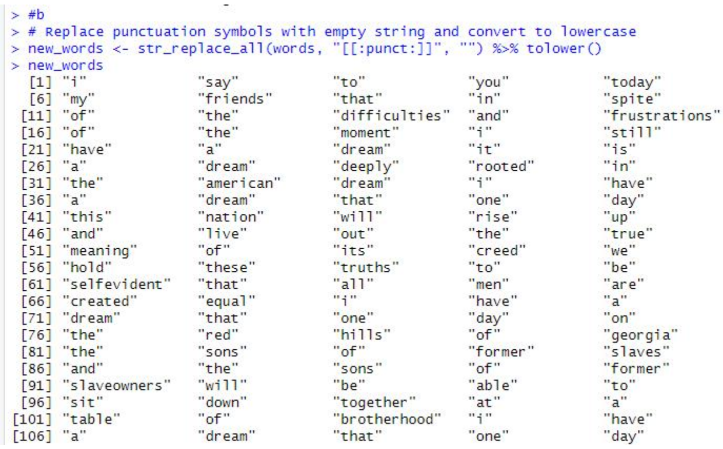


```
> print(punct_words)
 [1] "today,"         "friends,"       "moment,"        "dream."
 [5] "dream."         "creed:"         "self-evident:"  "equal."
 [9] "slave-owners"   "brotherhood."   "Mississippi,"   "state,"
[13] "oppression,"    "justice."       "character."     "today."
[17] "Alabama,"       "governor's"     "nullification," "brothers."
[21] "today."         "exalted,"       "low,"           "plain,"
[25] "straight,"      "revealed,"      "together."
```

B) :

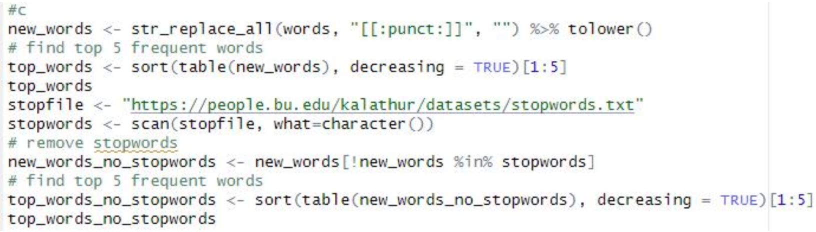


```
#b
# Replace punctuation symbols with empty string and convert to lowercase
new_words <- str_replace_all(words, "[[:punct:]]", "") %>% tolower()
new_words
```

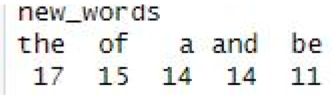


```
> #b
> # Replace punctuation symbols with empty string and convert to lowercase
> new_words <- str_replace_all(words, "[[:punct:]]", "") %>% tolower()
> new_words
  [1] "i"             "say"           "to"            "you"           "today"
  [6] "my"            "friends"       "that"          "in"            "spite"
 [11] "of"            "the"           "difficulties"  "and"           "frustrations"
 [16] "of"            "the"           "moment"        "i"             "still"
 [21] "have"          "a"             "dream"         "it"            "is"
 [26] "a"             "dream"         "deeply"        "rooted"        "in"
 [31] "the"           "american"      "dream"         "i"             "have"
 [36] "a"             "dream"         "that"          "one"           "day"
 [41] "this"          "nation"        "will"          "rise"          "up"
 [46] "and"           "live"          "out"           "the"           "true"
 [51] "meaning"       "of"            "its"           "creed"         "we"
 [56] "hold"          "these"         "truths"        "to"            "be"
 [61] "selfevident"   "that"          "all"           "men"           "are"
 [66] "created"       "equal"         "i"             "have"          "a"
 [71] "dream"         "that"          "one"           "day"           "on"
 [76] "the"           "red"           "hills"         "of"            "georgia"
 [81] "the"           "sons"          "of"            "former"        "slaves"
 [86] "and"           "the"           "sons"          "of"            "former"
 [91] "slaveowners"   "will"          "be"            "able"          "to"
 [96] "sit"           "down"          "together"      "at"            "a"
[101] "table"         "of"            "brotherhood"   "i"             "have"
[106] "a"             "dream"         "that"          "one"           "day"
```

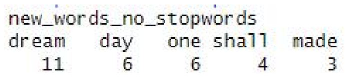
C) :



```
#c
new_words <- str_replace_all(words, "[[:punct:]]", "") %>% tolower()
# find top 5 frequent words
top_words <- sort(table(new_words), decreasing = TRUE)[1:5]
top_words
stopfile <- "https://people.bu.edu/kalathur/datasets/stopwords.txt"
stopwords <- scan(stopfile, what=character())
# remove stopwords
new_words_no_stopwords <- new_words[!new_words %in% stopwords]
# find top 5 frequent words
top_words_no_stopwords <- sort(table(new_words_no_stopwords), decreasing = TRUE)[1:5]
top_words_no_stopwords
```

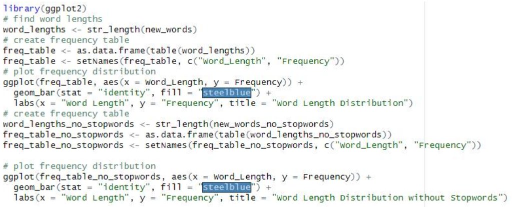


```
new_words
the  of   a and  be
 17  15  14  14  11
```



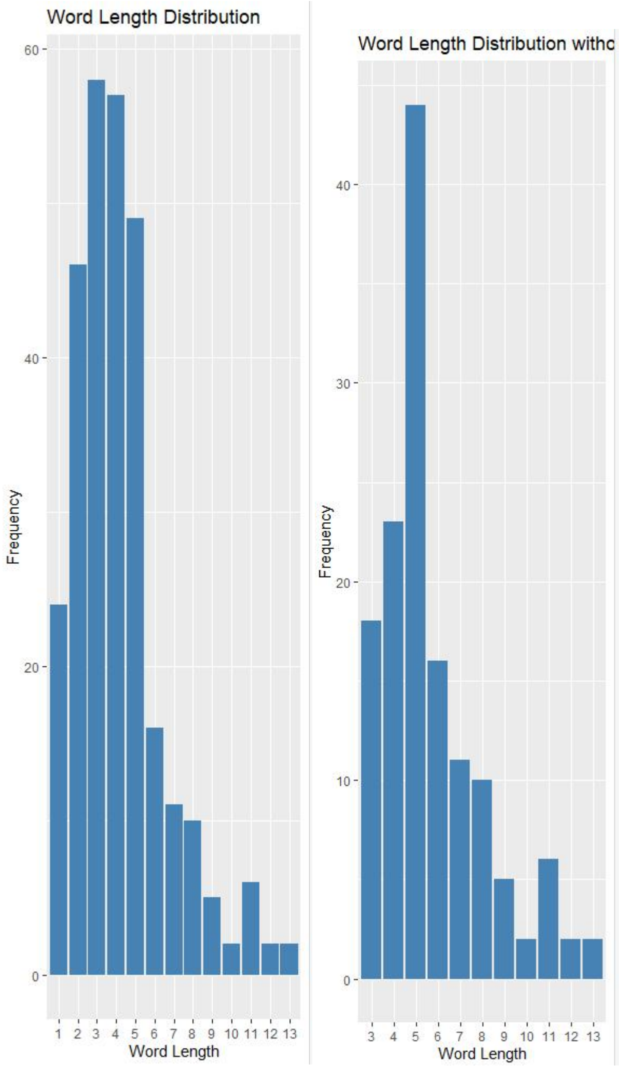
```
new_words_no_stopwords
dream   day   one shall  made
   11     6     6     4     3
```

D) :

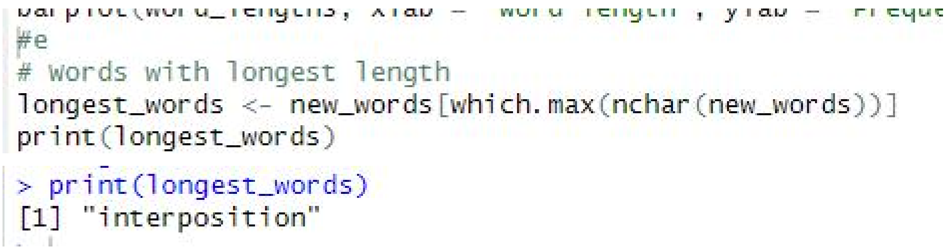


```
library(ggplot2)
# find word lengths
word_lengths <- str_length(new_words)
# create frequency table
freq_table <- as.data.frame(table(word_lengths))
freq_table <- setNames(freq_table, c("Word_Length", "Frequency"))
# plot frequency distribution
ggplot(freq_table, aes(x = Word_Length, y = Frequency)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(x = "Word Length", y = "Frequency", title = "Word Length Distribution")
# create frequency table
word_lengths_no_stopwords <- str_length(new_words_no_stopwords)
freq_table_no_stopwords <- as.data.frame(table(word_lengths_no_stopwords))
freq_table_no_stopwords <- setNames(freq_table_no_stopwords, c("Word_Length", "Frequency"))

# plot frequency distribution
ggplot(freq_table_no_stopwords, aes(x = Word_Length, y = Frequency)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(x = "Word Length", y = "Frequency", title = "Word Length Distribution without Stopwords")
```



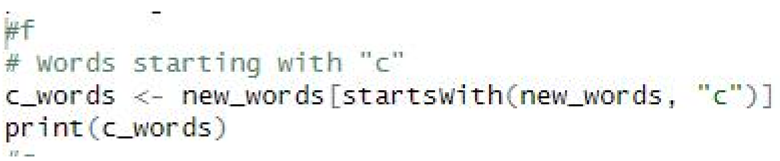
E) :



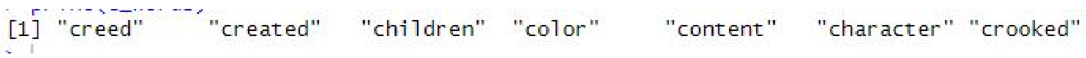
```
barplot(word_lengths, xlab = "word length", ylab = "Freque
#e
# Words with longest length
longest_words <- new_words[which.max(nchar(new_words))]
print(longest_words)
```

```
> print(longest_words)
[1] "interposition"
```

F) :

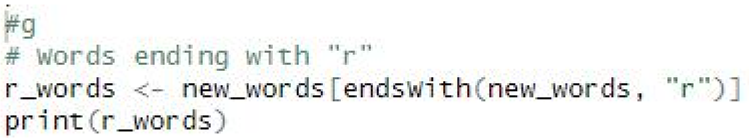


```
#f
# Words starting with "c"
c_words <- new_words[startsWith(new_words, "c")]
print(c_words)
```

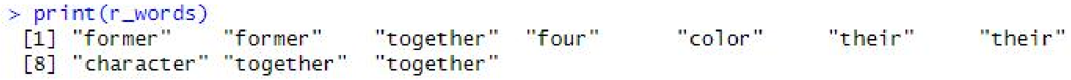


```
[1] "creed"     "created"   "children"  "color"     "content"   "character" "crooked"
```

G) :

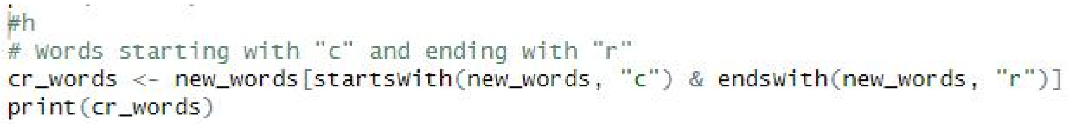


```
#g
# Words ending with "r"
r_words <- new_words[endsWith(new_words, "r")]
print(r_words)
```

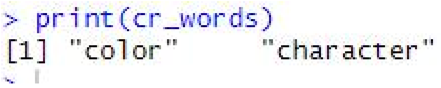


```
> print(r_words)
 [1] "former"    "former"    "together"  "four"      "color"     "their"     "their"    
 [8] "character" "together"  "together" 
```

H) :



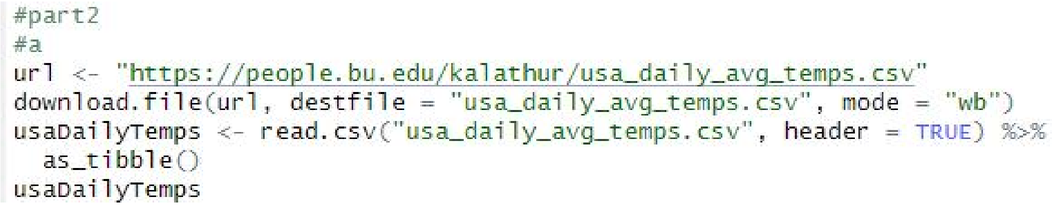
```
#h
# Words starting with "c" and ending with "r"
cr_words <- new_words[startsWith(new_words, "c") & endsWith(new_words, "r")]
print(cr_words)
```



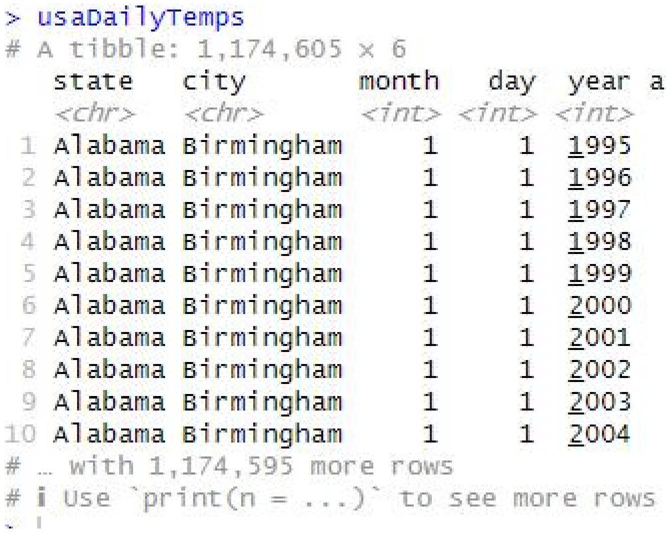
```
> print(cr_words)
[1] "color"     "character"
```

Part2

A) :

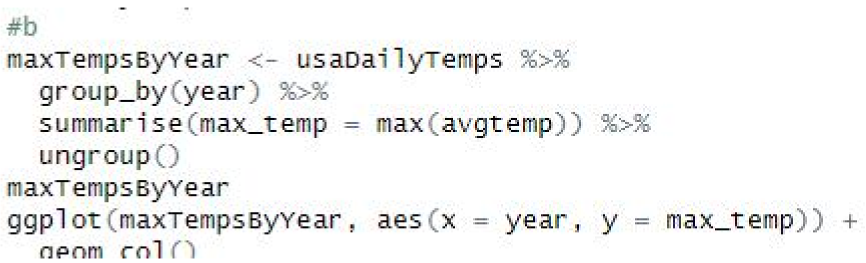


```
#part2
#a
url <- "https://people.bu.edu/kalathur/usa_daily_avg_temps.csv"
download.file(url, destfile = "usa_daily_avg_temps.csv", mode = "wb")
usaDailyTemps <- read.csv("usa_daily_avg_temps.csv", header = TRUE) %>%
  as_tibble()
usaDailyTemps
```

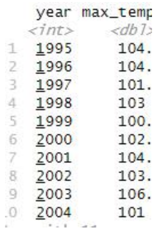


```
> usaDailyTemps
# A tibble: 1,174,605 × 6
   state   city       month   day  year avgtemp
   <chr>   <chr>      <int> <int> <int>   <dbl>
 1 Alabama Birmingham     1     1  1995    50.7
 2 Alabama Birmingham     1     1  1996    56.8
 3 Alabama Birmingham     1     1  1997    60.9
 4 Alabama Birmingham     1     1  1998    35.6
 5 Alabama Birmingham     1     1  1999    41  
 6 Alabama Birmingham     1     1  2000    59  
 7 Alabama Birmingham     1     1  2001    27  
 8 Alabama Birmingham     1     1  2002    28.1
 9 Alabama Birmingham     1     1  2003    51.7
10 Alabama Birmingham     1     1  2004    47.9
# … with 1,174,595 more rows
# ℹ Use `print(n = ...)` to see more rows
```

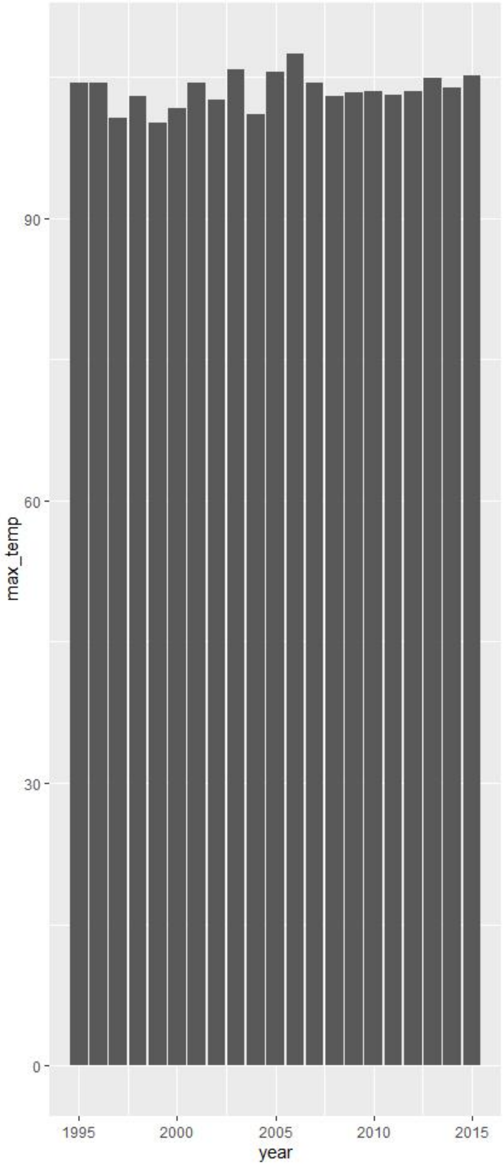
B) :



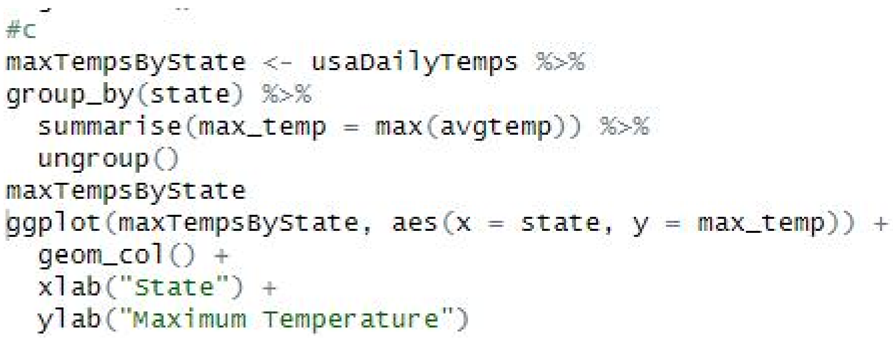
```
#b
maxTempsByYear <- usaDailyTemps %>%
  group_by(year) %>%
  summarise(max_temp = max(avgtemp)) %>%
  ungroup()
maxTempsByYear
ggplot(maxTempsByYear, aes(x = year, y = max_temp)) +
  geom_col()
```



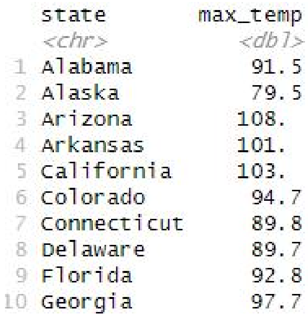
```
   year max_temp
  <int>    <dbl>
1  1995     104.
2  1996     104.
3  1997     101.
4  1998     103 
5  1999     100.
6  2000     102.
7  2001     104.
8  2002     103.
9  2003     106.
.0  2004     101 
```



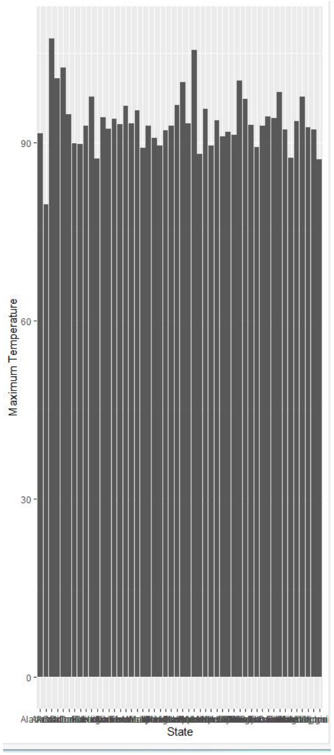
C) :



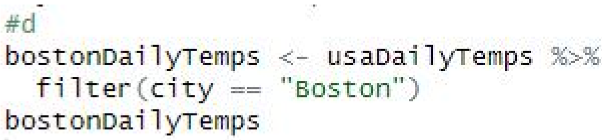
```
#c
maxTempsByState <- usaDailyTemps %>%
group_by(state) %>%
  summarise(max_temp = max(avgtemp)) %>%
  ungroup()
maxTempsByState
ggplot(maxTempsByState, aes(x = state, y = max_temp)) +
  geom_col() +
  xlab("State") +
  ylab("Maximum Temperature")
```



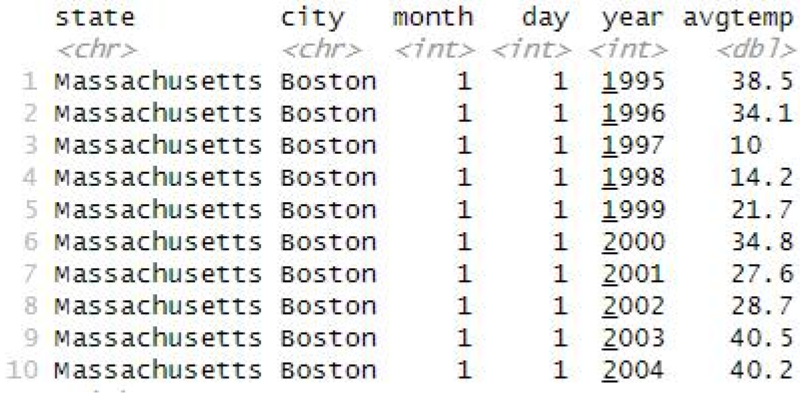
```
   state       max_temp
   <chr>          <dbl>
 1 Alabama         91.5
 2 Alaska          79.5
 3 Arizona        108.
 4 Arkansas       101.
 5 California     103.
 6 Colorado        94.7
 7 Connecticut     89.8
 8 Delaware        89.7
 9 Florida         92.8
10 Georgia         97.7
```



D) :

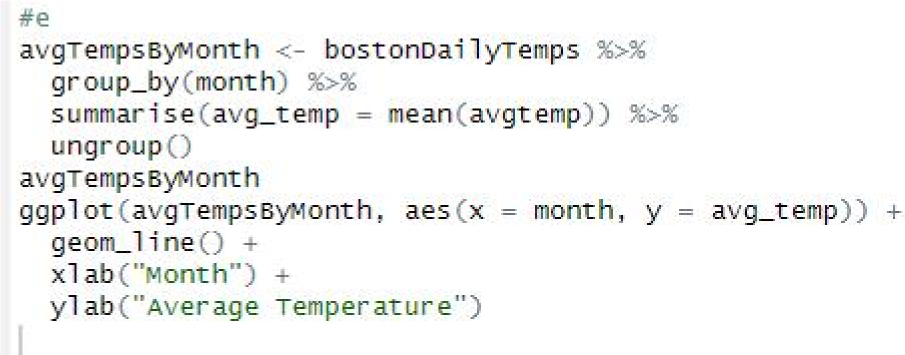


```
#d
bostonDailyTemps <- usaDailyTemps %>%
  filter(city == "Boston")
bostonDailyTemps
```

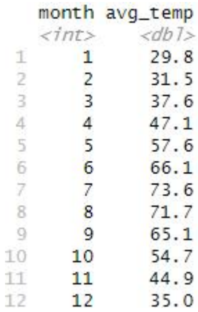


```
   state         city   month   day  year avgtemp
   <chr>         <chr>  <int> <int> <int>   <dbl>
 1 Massachusetts Boston     1     1  1995    38.5
 2 Massachusetts Boston     1     1  1996    34.1
 3 Massachusetts Boston     1     1  1997    10
 4 Massachusetts Boston     1     1  1998    14.2
 5 Massachusetts Boston     1     1  1999    21.7
 6 Massachusetts Boston     1     1  2000    34.8
 7 Massachusetts Boston     1     1  2001    27.6
 8 Massachusetts Boston     1     1  2002    28.7
 9 Massachusetts Boston     1     1  2003    40.5
10 Massachusetts Boston     1     1  2004    40.2
```

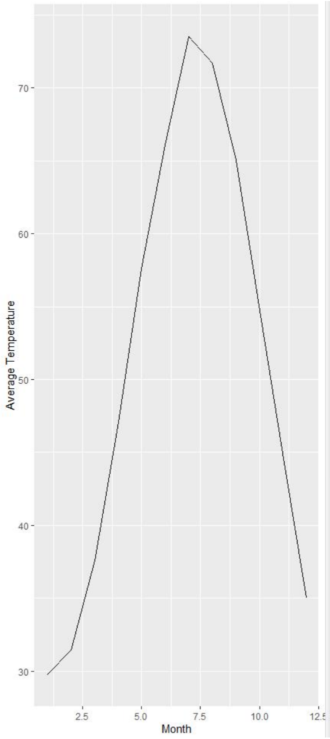
E) :



```
#e
avgTempsByMonth <- bostonDailyTemps %>%
  group_by(month) %>%
  summarise(avg_temp = mean(avgtemp)) %>%
  ungroup()
avgTempsByMonth
ggplot(avgTempsByMonth, aes(x = month, y = avg_temp)) +
  geom_line() +
  xlab("Month") +
  ylab("Average Temperature")
```



```
   month avg_temp
   <int>    <dbl>
 1     1     29.8
 2     2     31.5
 3     3     37.6
 4     4     47.1
 5     5     57.6
 6     6     66.1
 7     7     73.6
 8     8     71.7
 9     9     65.1
10    10     54.7
11    11     44.9
12    12     35.0
```


Average Temperature
Month