CS544, Fundamentals of Analytics Homework 6

Jefferson Parker

[japarker@bu.edu](mailto:japarker@bu.edu)

See the associated R code file HW6\_Parker.R for supporting code.

**Part 1, Strings with stringr**

1. Detect and show all the words with a punctuation symbol:

[1] "continent," "nation," "liberty," "equal." "war," "nation," "dedicated," "endure." "war."

[10] "field," "live." "this." "But," "sense," "ground." "men," "dead," "it,"

[19] "detract." "note," "here," "here." "living," "rather," "advanced." "nation," "God,"

[28] "people," "people," "people," "earth."

1. Replace the punctuation marks with an empty string and make this the new **words** data (just showing first five lines, see R code example for full content):

[1] "Four" "score" "and" "seven" "years" "ago" "our" "fathers"

[9] "brought" "forth" "upon" "this" "continent" "a" "new" "nation"

[17] "conceived" "in" "liberty" "and" "dedicated" "to" "the" "proposition"

[25] "that" "all" "men" "are" "created" "equal" "Now" "we"

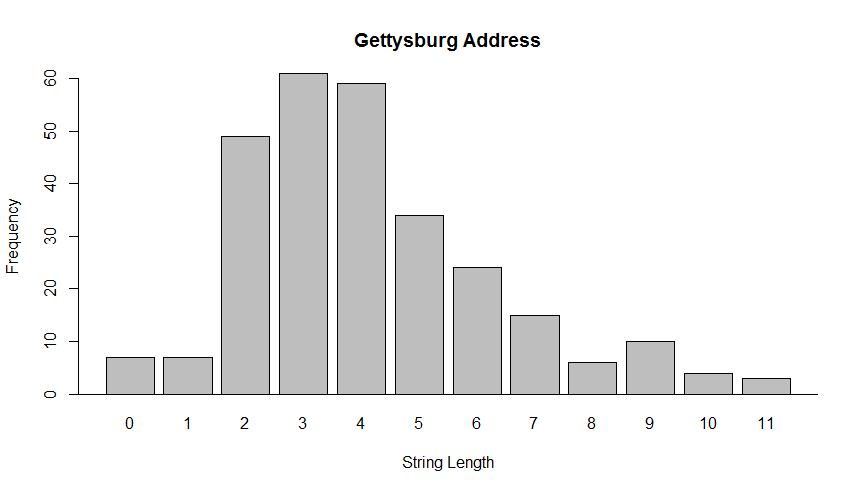
[33] "are" "engaged" "in" "a" "great" "civil" "war" "testing"

1. Show the frequencies and plot of the length of the new words data.

table(str\_length(words));

0 1 2 3 4 5 6 7 8 9 10 11

7 7 49 61 59 34 24 15 6 10 4 3



1. The words with longest length are:

words[str\_length(words) == max(str\_length(words))];

[1] "proposition" "battlefield" "consecrated"

with length = 11.

1. Show all words that start with the letter ‘p’:

[1] "proposition" "portion" "place" "proper" "poor" "power" "people" "people"

[9] "people" "perish"

1. Show all words that end with the letter ‘r’:

[1] "Four" "our" "war" "whether" "or" "war" "for" "their" "altogether"

[10] "proper" "larger" "far" "our" "poor" "power" "or" "nor" "remember"

[19] "never" "for" "rather" "far" "rather" "for" "for" "under" "for"

1. Show all words that start with ‘p’ AND end with ‘r’.

[1] "proper" "poor" "power"

**Part 2, Data Wrangling with tidyverse**

The data file was downloaded to a local directory and read into R using the read.csv function. For grading purposes, the file path in the code will have to be changed to the correct directory.

1. The data frame was converted to a tibble and named as directed:

> usaDailyTemps <- as\_tibble(tempdata);

> glimpse(usaDailyTemps);

Observations: 1,174,605

Variables: 6

$ state <fct> Alabama, Alabama, Alabama, Alabama, Alabama, Alabama,...

$ city <fct> Birmingham, Birmingham, Birmingham, Birmingham, Birmi...

$ month <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1 ...

$ day <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...

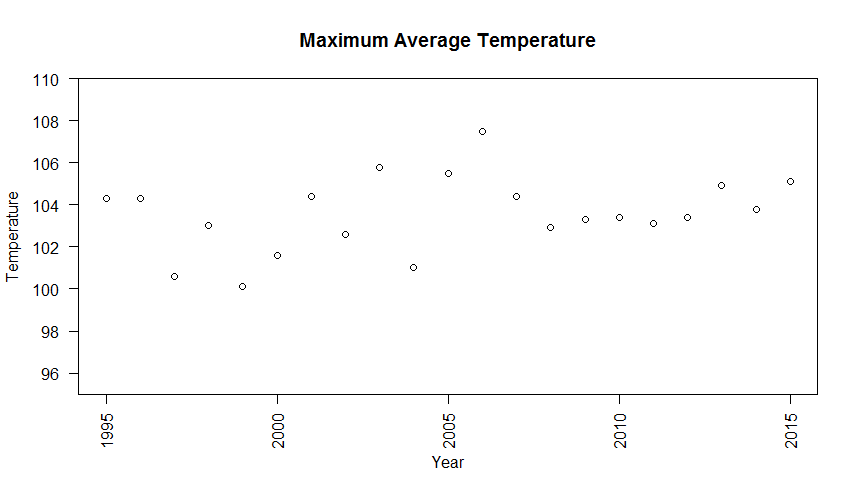
$ year <int> 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003,...

$ avgtemp <dbl> 50.7, 56.8, 60.9, 35.6, 41.0, 59.0, 27.0, 28.1, 51.7,...

1. The maximum temperatures recorded for each year were determined by grouping by the year parameter:

|  |  |
| --- | --- |
| Year | MaxTemp |
| 1995 | 104.3 |
| 1996 | 104.3 |
| 1997 | 100.6 |
| 1998 | 103 |
| 1999 | 100.1 |
| 2000 | 101.6 |
| 2001 | 104.4 |
| 2002 | 102.6 |
| 2003 | 105.8 |
| 2004 | 101 |
| 2005 | 105.5 |
| 2006 | 107.5 |
| 2007 | 104.4 |
| 2008 | 102.9 |
| 2009 | 103.3 |
| 2010 | 103.4 |
| 2011 | 103.1 |
| 2012 | 103.4 |
| 2013 | 104.9 |
| 2014 | 103.8 |
| 2015 | 105.1 |

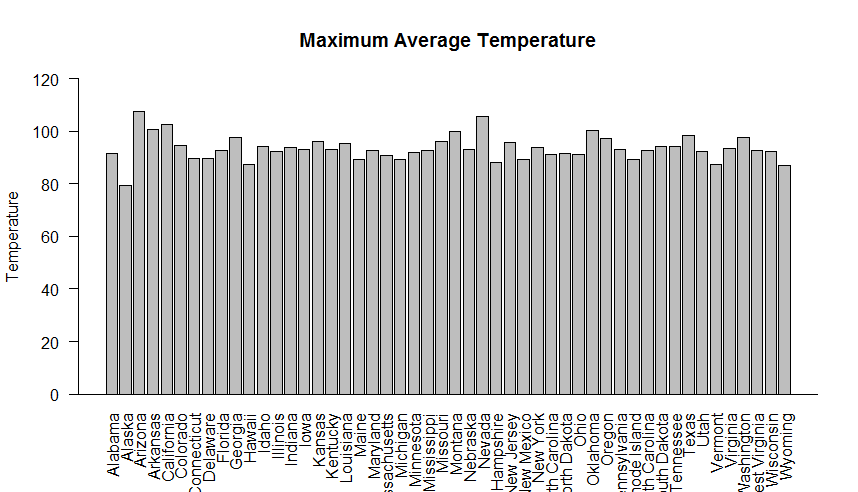
And plotted as:



1. The average temperatures recorded for each state were similarly determined by grouping on the state value:

|  |  |
| --- | --- |
| **State** | **MaxTemp** |
| Alabama | 91.5 |
| Alaska | 79.5 |
| Arizona | 107.5 |
| Arkansas | 100.7 |
| California | 102.6 |
| Colorado | 94.7 |
| Connecticut | 89.8 |
| Delaware | 89.7 |
| Florida | 92.8 |
| Georgia | 97.7 |
| Hawaii | 87.2 |
| Idaho | 94.2 |
| Illinois | 92.3 |
| Indiana | 94 |
| Iowa | 93 |
| Kansas | 96.1 |
| Kentucky | 93.2 |
| Louisiana | 95.4 |
| Maine | 89.1 |
| Maryland | 92.8 |
| Massachusetts | 90.7 |
| Michigan | 89.4 |
| Minnesota | 92 |
| Mississippi | 92.8 |
| Missouri | 96.3 |
| Montana | 100.1 |
| Nebraska | 93.2 |
| Nevada | 105.5 |
| New Hampshire | 88 |
| New Jersey | 95.6 |
| New Mexico | 89.4 |
| New York | 93.7 |
| North Carolina | 91 |
| North Dakota | 91.7 |
| Ohio | 91.2 |
| Oklahoma | 100.4 |
| Oregon | 97.3 |
| Pennsylvania | 92.9 |
| Rhode Island | 89.2 |
| South Carolina | 92.8 |
| South Dakota | 94.3 |
| Tennessee | 94.1 |
| Texas | 98.5 |
| Utah | 92.2 |
| Vermont | 87.4 |
| Virginia | 93.5 |
| Washington | 97.7 |
| West Virginia | 92.5 |
| Wisconsin | 92.2 |
| Wyoming | 87.1 |

And is plotted as



1. The data was filtered and assigned to a variable bostonDailyTemps using the filter function as:

bostonDailyTemps %>% group\_by(month) %>% summarise(meantemp = mean(avgtemp)) -> bostonMonthlyTemp;

1. The average monthly temperature in Boston is listed and plotted below:

|  |  |  |
| --- | --- | --- |
| **Month** | **MaxTemp** | **Month.Name** |
| 1 | 29.76667 | January |
| 2 | 31.47032 | February |
| 3 | 37.57604 | March |
| 4 | 47.08413 | April |
| 5 | 57.57803 | May |
| 6 | 66.10714 | June |
| 7 | 73.55038 | July |
| 8 | 71.68909 | August |
| 9 | 65.05762 | September |
| 10 | 54.73456 | October |
| 11 | 44.89366 | November |
| 12 | 34.99742 | December |

And plotted as:

