

DA5020 Homework 4: Strings and Factors

2018-02-12

Preparation

Download US [Farmers Markert Directory](#) data from the website of USDA (click on “Export to Excel”). Rename the file as *farmers_market.csv*.

Download the [Know Your Farmer, Know Your Food Projects](#) dataset and name it as *kyfprojects.xls*. Put it into the same folder.

Read the data:

```
library(tidyverse)
library(readxl)
farmers_dataset <- read_csv("farmers_market.csv")
kyfp_dataset <- read_excel("kyfprojects.xls")
```

Warm Up

This dataset stores city and state in different columns, what if you want to print out city and state in the format “City, State”?

Questions

Please edit this file and add your own solutions to these questions. Make your output as readable as possible. Next time you would need to create this file on your own. Feel free to try out other templates (e.g. [Tufte Handout](#)) if your are familiar with LaTeX. But for whatever template you choose, you should always include a link to your GitHub repo at the first page of your PDF.

1. (20 points) Cleanup the Facebook and Twitter column to let them contain only the facebook username or twitter handle name. I.e., replace
“<https://www.facebook.com/pages/Cameron-Park-Farmers-Market/97634216535?ref=hl>” with “Cameron-Park-Farmers-Market”,
“<https://twitter.com/FarmMarket125th>” with “FarmMarket125th”, and “@21acres” with “21acres”.

```
# read the data
farmers_dataset <- read_csv("farmers_market.csv")
farmers_dataset_fb <- as_vector(select(farmers_dataset, Facebook))
df <- data_frame(facebook = character())

# Facebook Column Clean
for(i in 1:NROW(farmers_dataset_fb))
{ x = farmers_dataset_fb[[i]]
  r1 = 'https://www.facebook.com/pages/'
```

```

r2 = 'https://www.facebook.com/'
r3 = 'www.facebook.com/pages/'
r4 = 'www.facebook.com/'
r5 = 'facebook.com/'
r6 = 'http://www.'
r7 = 'www.'
condition = if (grepl(r1,x)==TRUE){gsub(r1,"",x)}
else if(grepl(r2,x)==TRUE){gsub(r2,"",x)}
else if(grepl(r3,x)==TRUE){gsub(r3,"",x)}
else if(grepl(r4,x)==TRUE){gsub(r4,"",x)}
else if(grepl(r5,x)==TRUE){gsub(r5,"",x)}
else if(grepl(r6,x)==TRUE){gsub(r6,"",x)}
else if(grepl(r7,x)==TRUE){gsub(r7,"",x)}
else {x}
df[i, ] = condition}
df

## # A tibble: 8,710 x 1
##   facebook
##   * <chr>
## 1 Danville.VT.Farmers.Market/
## 2 StearnsHomesteadFarmersMarket
## 3 <NA>
## 4 <NA>
## 5 <NA>
## 6 12_South_Farmers_Market
## 7 125thStreetFarmersMarket
## 8 12th-Brandywine-Urban-Farm-Community-Garden/253769448091860
## 9 14UFarmersMarket
## 10 14KennedyFarmersMarket/
## # ... with 8,700 more rows

ans1_part1 <- data_frame(gsub("/.*", "", as.character(df$facebook)))

# Twitter Column Clean
farmers_twitter <- as_vector(select(farmers_dataset, Twitter))
df <- data_frame(twitter = character())
for(i in 1:NROW(farmers_twitter))
{ x = farmers_twitter[[i]]
  r1 = 'https://www.twitter.com/'
  r2 = 'https://twitter.com/'
  r3 = 'www.twitter.com'
  r4 = '/'
  r5 = 'twitter.com'
  r6 = '@'
  condition = if (grepl(r1,x) == TRUE){gsub(r1,"", x)}
  else if(grepl(r2,x) == TRUE){gsub(r2,"", x)}
  else if(grepl(r3,x) == TRUE){gsub(r3,"", x)}
  else if(grepl(r4,x) == TRUE){gsub(r4,"", x)}
  else if(grepl(r5,x) == TRUE){gsub(r5,"", x)}

```

```

else if(grepl(r6,x) == TRUE){gsub(r6,"", x)}
else {x}
df[i, ] = condition}
df

## # A tibble: 8,710 x 1
##   twitter
##   * <chr>
##   1 <NA>
##   2 <NA>
##   3 <NA>
##   4 <NA>
##   5 <NA>
##   6 12southfrmsmkt
##   7 FarmMarket125th
##   8 <NA>
##   9 14UFarmersMkt
##  10 14KenFM
## # ... with 8,700 more rows

ans1_part2 <- df

```

2. (20 points) Clean up the city and street column. Remove state and county names from the city column and consolidate address spellings to be more consistent (e.g. "St.", "ST.", "Street" all become "St"; "and" changes to "&", etc...).

```

#read the data
farmers_dataset <- read_csv("farmers_market.csv.csv")
# to consolidate address spellings to be more consistent
farmers_dataset$street<-gsub(" Street", "
St",as.character(farmers_dataset$street),fixed=TRUE)
farmers_dataset$street<-gsub(" street", "
St",as.character(farmers_dataset$street),fixed=TRUE)
farmers_dataset$street<-gsub(" and ", " &
",as.character(farmers_dataset$street),fixed=TRUE)
farmers_dataset$street<-gsub(" ST.", "
St",as.character(farmers_dataset$street),fixed=TRUE)
farmers_dataset$street<-gsub(" St.", "
St",as.character(farmers_dataset$street),fixed=TRUE)
farmers_dataset$street<-gsub(" Sts", "
St",as.character(farmers_dataset$street),fixed=TRUE)
farmers_dataset$street<-gsub(" Boulevard", "
Blvd",as.character(farmers_dataset$street),fixed=TRUE)
# to consolidate city
farmers_dataset$city <-
gsub(",.*", "",as.character(farmers_dataset$city),fixed=FALSE)

```

3. (20 points) Create a new data frame (tibble) that explains the online presence of each state's farmers market. I.e., how many percentages of them have a facebook account? A twitter account? Or either of the accounts? (Hint: use the is.na() function)

```
library("tibble")
#read the data
farmers_dataset <- read_csv("farmers_market.csv.csv")
#create the table for online presense
Table1 <- farmers_dataset %>%
  group_by(State) %>%
  summarise( Facebook = ((sum(!is.na(Facebook)))/ n())*100,
             twitter = ((sum(!is.na(Twitter)))/ n())*100,
             Youtube = ((sum(!is.na(Youtube)))/n())*100,
             Other_media = ((sum(!is.na(OtherMedia)))/n())*100)
```

4. (20 points) Some of the farmer market names are quite long. Can you make them shorter by using the `forcats::fct_recode` function? Create a plot that demonstrates the number of farmers markets per location type. The locations should be ordered in descending order where the top of the graph will have the one with the highest number of markets.

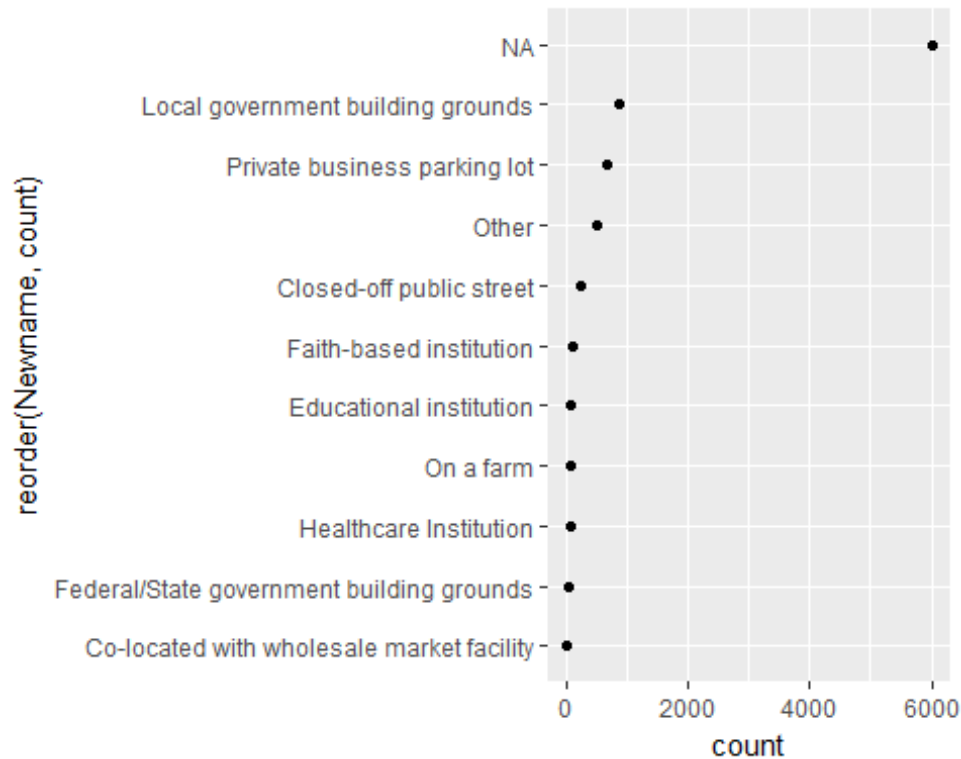
```
#read the data
farmers_dataset <- read_csv("farmers_market.csv.csv")

#Create the plot with shorten names on the axis

plot_data <- farmers_dataset %>%
  group_by(Location) %>%
  summarise(count = n()) %>%
  as_tibble() %>%

  mutate(Newname = recode_factor(Location, `On a farm from: a barn, a
greenhouse, a tent, a stand, etc` = "On a farm", `Faith-based institution
(e.g., church, mosque, synagogue, temple)` = "Faith-based institution"))

ggplot(data = plot_data, mapping = aes(y = reorder(Newname, count) , x =
count) )+ geom_point()
```



5. (20 points) Write code to sanity check the kyfprojects data. For example, does Program Abbreviation always match Program Name for all the rows? (Try thinking of your own rules, too.)

```
kyfp <- read_excel("kyfprojects.xls")

#Zipcode check if consistent in format
zip2 <- grepl("^[0-9]+$", select(kyfp, Zip))
zip2

## [1] FALSE

# Check State characters
state_data <- nchar(kyfp$State)
sum(state_data/2)

## [1] 2379

table(state_data)["2"]

##      2
## 2379

# Check states names
as_data_frame(unique(kyfp$State))

## # A tibble: 56 x 1
##   value
##   <chr>
```

```
## 1 IL
## 2 MN
## 3 NM
## 4 LA
## 5 AZ
## 6 WV
## 7 AL
## 8 DC
## 9 NJ
## 10 ND
## # ... with 46 more rows
```

```
# Check year
```

```
year <- grepl("(20[09]\\d|20[12]\\d)", kyfp$Year)
year
```

```
## [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [14] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [27] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [40] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [53] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [66] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [79] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [92] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [105] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [118] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [131] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [144] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [157] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [170] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [183] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [196] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [209] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [222] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [235] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [248] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [261] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [274] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [287] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [300] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [313] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [326] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [339] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [352] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [365] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [378] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [391] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [404] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [417] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [430] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

[illegible]

[illegible]

[illegible]

```
# Check recipient type
```

```
Unique_recipient <- distinct(select(kyfp, 'Recipient Type'))
```

```
Unique_recipient
```

```
## # A tibble: 8 x 1
##   `Recipient Type`
##   <chr>
## 1 Business
## 2 Nonprofit
## 3 Government
## 4 Academic
## 5 Other
## 6 Producer
## 7 Nonprofit Academic
## 8 Businesses
```

```
# Check Funding type
```

```
fund <- distinct(select(kyfp, 'Funding Type'))
```

```
fund
```

```
## # A tibble: 2 x 1
##   `Funding Type`
##   <chr>
## 1 Grant
## 2 Loan
```

```
# To Check if town is char
```

```
town <- grepl("[A-z]", kyfp$Town)
```

```
town
```

```
##   [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##   [14] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##   [27] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##   [40] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##   [53] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##   [66] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##   [79] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##   [92] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [105] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [118] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [131] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [144] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [157] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [170] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [183] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [196] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [209] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [222] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [235] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [248] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [261] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

[illegible]

[illegible]

[illegible]

```
## [2224] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2237] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2250] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2263] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2276] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2289] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2302] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2315] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2328] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2341] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2354] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2367] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

Submission

You need to submit an .Rmd extension file as well as the generated pdf file. Be sure to state all the assumptions and give explanations as comments in the .Rmd file wherever needed to help us assess your submission. Please name the submission file LAST_FirstInitial_1.Rmd for example for John Smith's 1st assignment, the file should be named Smith_J_1.Rmd.