Class Activity 9

Arthur Viegas Eguia

April 12 2024

Problem 1

a) Use `read_csv()` to import the `desserts` data set from GitHub url. Use `glimpse` to see if the data import is alright.

```
url <- "https://raw.githubusercontent.com/deepbas/statdatasets/main/desserts.csv"
desserts <- readr::read_csv(url)
glimpse(desserts)
Rows: 549
Columns: 16
$ series                <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
$ episode               <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, ~
$ baker                 <chr> "Annetha", "David", "Edd", "Jasminder", "Jonatha~
$ technical             <chr> "2nd", "3rd", "1st", "N/A", "9th", "N/A", "8th",~
$ result                <chr> "IN", "IN", "IN", "IN", "IN", "IN", "IN", "IN", ~
$ uk_airdate            <chr> "17 August 2010", "17 August 2010", "17 August 2~
$ us_season             <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
$ us_airdate            <date> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ showstopper_chocolate <chr> "chocolate", "chocolate", "no chocolate", "no ch~
$ showstopper_dessert   <chr> "other", "other", "other", "other", "other", "ca~
$ showstopper_fruit     <chr> "no fruit", "no fruit", "no fruit", "no fruit", ~
$ showstopper_nut       <chr> "no nut", "no nut", "no nut", "no nut", "almond"~
$ signature_chocolate   <chr> "no chocolate", "chocolate", "no chocolate", "no~
$ signature_dessert     <chr> "cake", "cake", "cake", "cake", "cake", "cake", ~
$ signature_fruit       <chr> "no fruit", "fruit", "fruit", "fruit", "fruit", ~
$ signature_nut         <chr> "no nut", "no nut", "no nut", "no nut", "no nut"~
```

Does everything look good? Import the dataset with correct data types, if needed. Fix the problems, if any.

```
# your r-code
desserts1 <- read_csv(url, col_types = list(
    technical=col_number(),
    uk_airdate = col_date()
  ))
glimpse(desserts1)
Rows: 549
Columns: 16
```

```
$ series                <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
$ episode               <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, ~
$ baker                 <chr> "Annetha", "David", "Edd", "Jasminder", "Jonatha~
$ technical             <dbl> 2, 3, 1, NA, 9, NA, 8, NA, 10, NA, 8, 6, 2, 1, 3~
$ result                <chr> "IN", "IN", "IN", "IN", "IN", "IN", "IN", "IN", ~
$ uk_airdate            <date> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ us_season             <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
$ us_airdate            <date> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ showstopper_chocolate <chr> "chocolate", "chocolate", "no chocolate", "no ch~
$ showstopper_dessert   <chr> "other", "other", "other", "other", "other", "ca~
$ showstopper_fruit     <chr> "no fruit", "no fruit", "no fruit", "no fruit", ~
$ showstopper_nut       <chr> "no nut", "no nut", "no nut", "no nut", "almond"~
$ signature_chocolate   <chr> "no chocolate", "chocolate", "no chocolate", "no~
$ signature_dessert     <chr> "cake", "cake", "cake", "cake", "cake", "cake", ~
$ signature_fruit       <chr> "no fruit", "fruit", "fruit", "fruit", "fruit", ~
$ signature_nut         <chr> "no nut", "no nut", "no nut", "no nut", "no nut"~
problems(desserts1)
# A tibble: 556 x 5
     row   col expected        actual         file
   <int> <int> <chr>           <chr>          <chr>
 1     2     6 date in ISO8601 17 August 2010 ""
 2     3     6 date in ISO8601 17 August 2010 ""
 3     4     6 date in ISO8601 17 August 2010 ""
 4     5     4 a number        N/A            ""
 5     5     6 date in ISO8601 17 August 2010 ""
 6     6     6 date in ISO8601 17 August 2010 ""
 7     7     4 a number        N/A            ""
 8     7     6 date in ISO8601 17 August 2010 ""
 9     8     6 date in ISO8601 17 August 2010 ""
10     9     4 a number        N/A            ""
# i 546 more rows
```

```
desserts2 <- read_csv(url, col_types = list(
  technical = col_number(),
  uk_airdate = col_date(format = "%d %B %Y")
))
problems(desserts2)
# A tibble: 7 x 5
    row   col expected actual file
  <int> <int> <chr>    <chr>  <chr>
1     5     4 a number N/A    ""
2     7     4 a number N/A    ""
3     9     4 a number N/A    ""
4    11     4 a number N/A    ""
5    35     4 a number N/A    ""
6    36     4 a number N/A    ""
7    37     4 a number N/A    ""
```

```
desserts3 <- read_csv(url, col_types = list(
  technical = col_number(),
  uk_airdate = col_date(format = "%d %B %Y")
),
na = c("N/A", "", "unknown", "NA")
)
```

```
problems(desserts3)
# A tibble: 0 x 5
# i 5 variables: row <int>, col <int>, expected <chr>, actual <chr>, file <chr>
```

Problem 2

a) Use the appropriate `read_<type>()` function to import the datasets provided in the course webpage. Identify and fix any issues you encounter.

If you hit any errors/problems, be sure to explore them and identify the issue, even if you can't "fix" it.

```
# your code for simple 1
simple <- readr::read_csv("data/simple-1.dat")
#write_csv(simple, "data/simple_copy.csv")
```

```
# your code for mild 1
read_csv("data/mild-1.csv")
# A tibble: 10 x 1
   `first|last|address|city|postcode`
   <chr>
 1 Leah|Downs|688-5741 Ut St.|Owensboro|V9Z 9K2
 2 Boris|Kirby|257-5422 Vel Avenue|Rialto|C6I 9S0
 3 Naida|Franco|809-5528 Tristique Avenue|Atwater|T8K 7U8
 4 Xena|Tucker|7218 A St.|Grand Forks|M6O 1X4
 5 Rylee|Wise|155-6070 Purus. St.|Bradford|65359
 6 Baxter|Gallagher|2415 Ligula. St.|Carbondale|55211
 7 Griffin|Benjamin|3261 Ac St.|Guayama|94450
 8 Rinah|Bradley|787-9626 Eget Avenue|Norton|17673
 9 Tobias|Walter|4717 Mauris. Street|Attleboro|73678
10 Boris|Farley|893-8193 Quisque Avenue|San Clemente|74492
mild1 <- read_delim("data/mild-1.csv")
```

```
# your code for tricky 1
tricky1 <- read_csv("data/tricky-1.csv")
problems(tricky1)
# A tibble: 2 x 5
    row   col expected  actual    file
  <int> <int> <chr>     <chr>     <chr>
1     4     4 5 columns 4 columns /Accounts/viegaseguiaa/stat220/Class activity~
2     7     4 5 columns 4 columns /Accounts/viegaseguiaa/stat220/Class activity~
```

The issue is that we have missing values that aren't specifically included in the rows 4 and 7 of the **original** file (so rows 3 and 6 once we load the data). We can fix this with post processing

```
# your post-processing code
tricky1[3, ] <- c(tricky1[3, 1:2], NA, tricky1[3, 3:4])
tricky1[6, ] <- c(tricky1[6, 1], NA, tricky1[6, 2:4])

tricky1
# A tibble: 10 x 5
   first     last     address                city         postcode
   <chr>     <chr>    <chr>                  <chr>        <chr>
 1 Leah      Downs    688-5741 Ut St.        Owensboro    V9Z 9K2
 2 Boris     Kirby    257-5422 Vel Avenue    Rialto       C6I 9S0
 3 Naida     Franco   <NA>                   Atwater      T8K 7U8
 4 Xena      Tucker   7218 A St.             Grand Forks  M6O 1X4
```

```
 5 Rylee     Wise     155-6070 Purus. St.     Bradford     65359
 6 Gallagher <NA>     2415 Ligula. St.        Carbondale   55211
 7 Griffin   Benjamin 3261 Ac St.             Guayama      94450
 8 Rinah     Bradley  787-9626 Eget Avenue    Norton       17673
 9 Tobias    Walter   4717 Mauris. Street     Attleboro    73678
10 Boris     Farley   893-8193 Quisque Avenue San Clemente 74492
```

```
# your code for tricky 2
tricky2 <- read_csv("data/tricky-2.csv")
problems(tricky2)
# A tibble: 0 x 5
# i 5 variables: row <int>, col <int>, expected <chr>, actual <chr>, file <chr>
tricky2
# A tibble: 16 x 6
   iata  airport                  city                 latitude longitude   ...6
   <chr> <chr>                    <chr>                <chr>        <dbl>  <dbl>
 1 00M   Thigpen                  Bay Springs, MS      31.9537~     -89.2   NA
 2 00R   Livingston Municipal     Livingston, TX       30.6858~     -95.0   NA
 3 00V   Meadow Lake              Colorado Springs, CO 38.9457~    -105.    NA
 4 01G   Perry-Warsaw             Perry, NY            42.7413~     -78.1   NA
 5 01J   Hilliard Airpark         Hilliard, FL         30.6880~     -81.9   NA
 6 01M   Tishomingo County        Belmont, MS          34.4916~     -88.2   NA
 7 02A   Gragg-Wade               Clanton, AL          32.8504~     -86.6   NA
 8 02C   Capitol                  Brookfield           WI            43.1  -88.2
 9 02G   Columbiana County        East Liverpool       OH            40.7  -80.6
10 03D   Memphis Memorial         Memphis              MO            40.4  -92.2
11 04M   Calhoun County           Pittsboro            MS            33.9  -89.3
12 04Y   Hawley Municipal         Hawley               MN            46.9  -96.4
13 05C   Griffith-Merrillville    Griffith             IN            41.5  -87.4
14 05F   Gatesville - City/County Gatesville           TX            31.4  -97.8
15 05U   Eureka                   Eureka               NV            39.6 -116.
16 07M   Riverside Municipal      Riverside            CA            34.0 -117.
```

Why are there state abbreviations in the latitude column?

```
# your post-processing code
tricky2_part1 <- read_csv("data/tricky-2.csv", n_max = 7)



tricky2_part1 <- tricky2_part1 %>% separate(city, c("city, state"), sep=", ")
names <- colnames(tricky2_part1)
tricky2_part1

tricky2_part2 <- read_csv("data/tricky-2.csv", skip = 8, col_names = names)
full_join(tricky2_part1, tricky2_part2) -> fullCleanData
```

Acknowledgemant

Parts of the activities are adapted from similar activity written by Adam Loy.