HW3: Tidyverse Work

Mike Maccia

Loading libraries

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.2     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.4
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```
library(palmerpenguins)
```

```

Attaching package: 'palmerpenguins'

The following objects are masked from 'package:datasets':

    penguins, penguins_raw
```

Task 1

Question a: Why `read_csv` cannot be used to read the data.txt file

The function `read_csv()` can only be used to read in files that use the delimeters of commas or tabs. The function `read_csv2()` must be used in files with semicolons as the separator (commas can be used for decimal points).

```
data <- read.csv2('~/ST558 Repo/HW3_/Data/data.txt',
                  header = T)
data
```

```
  x y z
1 1 2 3
2 5 3 8
```

Question b: Reading in 2nd file

In this file, “6” is the delimeter.

```
data_2 <- read_delim('~/ST558 Repo/HW3_/Data/data2.txt',
                     delim = '6',
                     col_types= 'fdc')

data_2
```

```
# A tibble: 3 x 3
  x         y z
  <fct> <dbl> <chr>
1 1         2 3
2 5         3 8
3 7         4 2
```

Task 2

Data tidying skills

Question a: Reading Data

Reading in the `trailblazer.csv` data

```
trailblazer <- read_csv('~/ST558 Repo/HW3_/Data/trailblazer.csv',
                   col_names= TRUE)
```

```
Rows: 9 Columns: 11
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (1): Player
dbl (10): Game1_Home, Game2_Home, Game3_Away, Game4_Home, Game5_Home, Game6_...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
trailblazer
```

```
# A tibble: 9 x 11
  Player       Game1_Home Game2_Home Game3_Away Game4_Home Game5_Home Game6_Away
  <chr>             <dbl>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>
1 Damian Lill~         20         19         12         20         25         14
2 CJ McCollum          24         28         20         25         14         25
3 Norman Powe~         14         16         NA         NA         12         14
4 Robert Covi~          8          6          0          3          9          6
5 Jusuf Nurkic         20          9          4         17         14         13
6 Cody Zeller           5          5          8         10          9          6
7 Anfernee Si~         11         18         12         17          5         19
8 Larry Nance~          2          8          5          8          3          8
9 Nassir Litt~          7         11          5          9          8          8
# i 4 more variables: Game7_Away <dbl>, Game8_Away <dbl>, Game9_Home <dbl>,
#   Game10_Home <dbl>
```

Question b: Pivoting the data longer

```
trailblazer_longer <- trailblazer |>
    pivot_longer(cols = 2:11,
                 names_to = c('Game', 'Location'),
                 names_prefix = 'Game',
                 names_sep = '_',
                 values_to = 'Points')

print(trailblazer_longer, n=5)
```

```
# A tibble: 90 x 4
  Player         Game  Location Points
  <chr>          <chr> <chr>     <dbl>
```

```
1 Damian Lillard 1     Home          20
2 Damian Lillard 2     Home          19
3 Damian Lillard 3     Away          12
4 Damian Lillard 4     Home          20
5 Damian Lillard 5     Home          25
# i 85 more rows
```

Question c: Who scored more when playing at home versus away

```
library(kableExtra)
```

```

Attaching package: 'kableExtra'

The following object is masked from 'package:dplyr':

    group_rows
```

```
trailblazer_wider <- trailblazer_longer |>
    pivot_wider(names_from = (Location),
                values_from = Points) |> #pivot wider
    group_by(Player) |>
    summarize(mean_home = mean(Home, na.rm=T), #creating mean of home points/away points
                                                #by player
              mean_away = mean (Away, na.rm=T)) |>
    mutate(diff_points = (mean_home - mean_away)) |> #creating different in means
    arrange(desc(diff_points)) |>#arrange by descending
    kbl(digits=2)

trailblazer_wider
```

Player	mean_home	mean_away	diff_points
Jusuf Nurkic	14.17	7.50	6.67
Robert Covington	9.50	3.00	6.50
Nassir Little	8.33	4.25	4.08
Damian Lillard	18.83	18.00	0.83
Cody Zeller	5.83	5.25	0.58
Larry Nance Jr	4.50	5.00	-0.50
CJ McCollum	20.83	21.50	-0.67
Anfernee Simons	12.83	15.75	-2.92
Norman Powell	16.00	19.67	-3.67

While they did not necessarily score the most points, Jusuf Nurkic (6.67) and Robert Covington (6.5) scored on average more points at home than away through the first 10 games of the season.

Task 3

Question a. Describing what some values mean

indicates that there were no values within a column. For example, there were no bill_length measurements for Gentoo species on Torgersen island.

<dbl [52]> indicates that within that cell there would are 52 observations (which are doubles) for bill length.

indicates a list-column within a tribble. List-columns occur when each element within a column is a list. For example, the above cell of Adelie species on Torgersen island, there is a list of 52 doubles within that cell / element.

Question b. Creating a new table

```r
penguins_island_ct <- penguins |>
  select(species, island)  |>
  group_by(species, island) |>
  summarize(count=n(), .groups='drop') |>
  pivot_wider(names_from = island,
              values_from = count,
              values_fill = 0)

penguins_island_ct
```

```
# A tibble: 3 x 4
  species   Biscoe Dream Torgersen
  <fct>      <int> <int>     <int>
1 Adelie        44    56        52
2 Chinstrap      0    68         0
3 Gentoo       124     0         0
```

Task 4

Replacing 2 missing values for bill length

```
penguins_fixed_bill_length <- penguins |>
  mutate(bill_length_mm =
           case_when(species == "Gentoo" & is.na(bill_length_mm)
                                 ~ 30, TRUE ~ bill_length_mm )) |>
                  #finding where gentoo is missing change to 30
  mutate(bill_length_mm =
           case_when(species == "Adelie" & is.na(bill_length_mm)
                                 ~ 26, TRUE ~ bill_length_mm )) |>
                  #finding where adelie is missing change to 26
  arrange(bill_length_mm)

print(penguins_fixed_bill_length, n = 10)
```

```
# A tibble: 344 x 8
   species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
   <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
 1 Adelie  Torgersen           26            NA                  NA          NA
 2 Gentoo  Biscoe              30            NA                  NA          NA
 3 Adelie  Dream               32.1          15.5               188        3050
 4 Adelie  Dream               33.1          16.1               178        2900
 5 Adelie  Torgersen           33.5          19                 190        3600
 6 Adelie  Dream               34            17.1               185        3400
 7 Adelie  Torgersen           34.1          18.1               193        3475
 8 Adelie  Torgersen           34.4          18.4               184        3325
 9 Adelie  Biscoe              34.5          18.1               187        2900
10 Adelie  Torgersen           34.6          21.1               198        4400
# i 334 more rows
# i 2 more variables: sex <fct>, year <int>
```