Tidyverse Create - Extended

Coco Donovan, Kory Martin

2023-04-29

Introduction:

For my tidyverse create assignment, I chose a data set containing roster information for all NCAA Women's Basketball teams. I intend to use readr to read in my data, dplyr to manipulate my data and ggplot2 to display my analysis.

Loading/Installing Packages

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(readr)
```

readr

read_csv()

I used the read_csv() function to read in my csv of NCAA Women's Basketball roster info. Within read_csv() I used the col_types argument to change data types of certain columns. Initially the columns that I changed inside read_csv() were chr type, but to do meaningful analysis I needed to make them numerical, hence setting them to col_double() (which sets the data type of the column to a double). This same logic can be applied to height_

```
ncaa_wbb_rosters <- read_csv('https://raw.githubusercontent.com/Sports-Roster-Data/womens-college-baske

ncaa_avg_height <- round(mean(ncaa_wbb_rosters$total_inches, na.rm = TRUE),2)
```

dyplr

distinct()

I am unsure what values the redshirt column could possibly take on, so I use a pipe and the distinct() function to highlight the possible values 'redshirt' may take on. I find that redshirt can either be 1, for "yes, a student athlete was redshirted," or 0, for "no, a student athlete was not redshirted."

```
knitr::kable(ncaa_wbb_rosters %>%
  distinct(redshirt))
```

redshirt
0
1

select(), group_by(), summarize(), and arrange()

Now, I wanted to get a glimpse of the teams with the tallest average height. To do this I used pipes. I selected the team variable and the total_inches variable. Then I grouped by height using the group_by function. I then used the summarize function to provide the counts of players for players per team the corresponding average height per team.

```
avg_heights <- ncaa_wbb_rosters %>%
  select('team','total_inches') %>%
  group_by(team) %>%
  summarize(number_of_players = n(), Avg_height = round(mean(total_inches),2)) %>%
  arrange(desc(Avg_height))

knitr::kable(head(avg_heights))
```

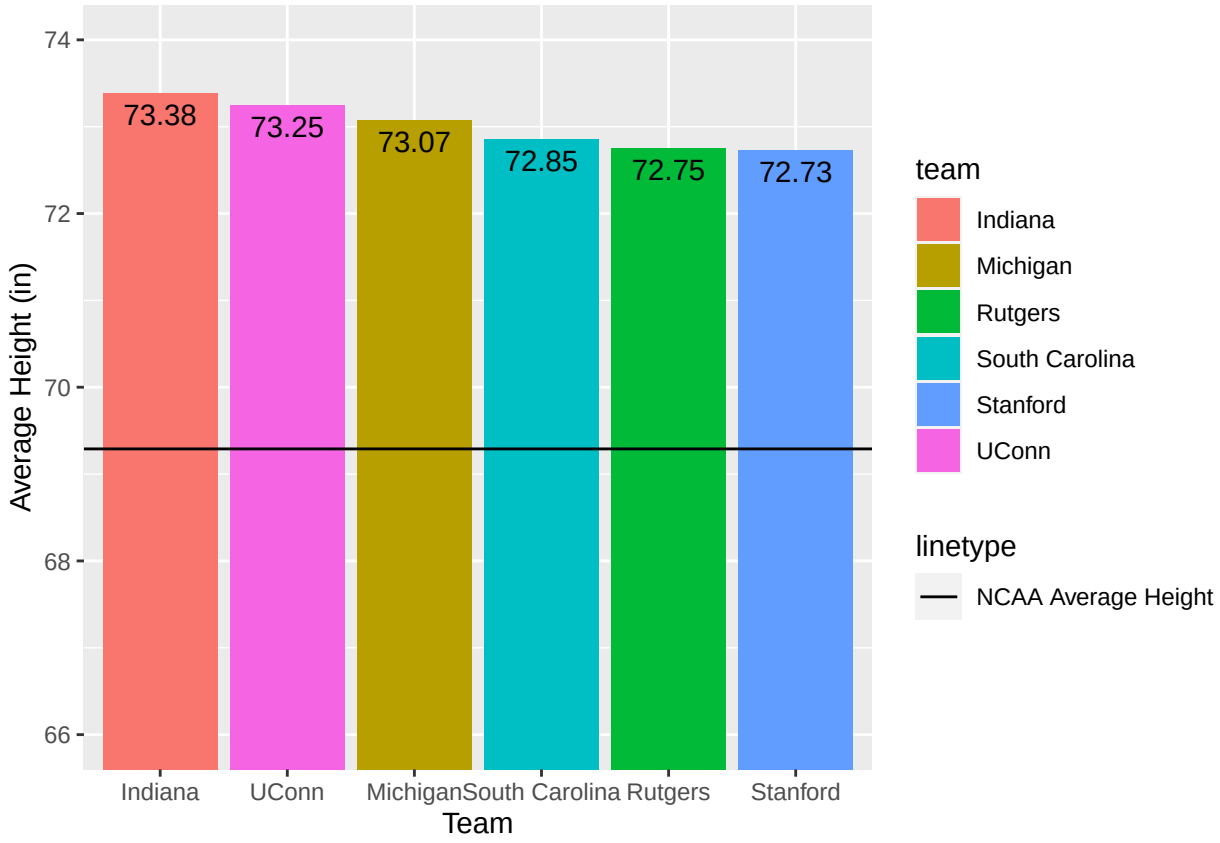
team	number_of_players	Avg_height
Indiana	13	73.38
UConn	12	73.25
Michigan	15	73.07
South Carolina	13	72.85
Rutgers	8	72.75
Stanford	15	72.73

ggplot2

Plotting Average Heights per team

The primary package I used here was ggplot2, although I do start this code chunk off with some dplyr. I used top_n() so that I could get the stop six results for the teams with the highest average heights. Then I piped those top 6 results into a bar chart. I made sure to change the order of the x-axis as the default is to sort the items alphabetically, but I want the value to be sorted based height in a descending order. I then limited the x-axis to better range to show the difference between the top average heights (using coord_cartesian()), and finished off my visualization by displaying the numerical values on top of each bar using geom_text() and displaying a line to show the average height of an NCAA women's basketball team for perspective.

```
top_n(avg_heights, n=6, Avg_height) %>%
  ggplot(., mapping = aes(x=reorder(team, desc(Avg_height)),y=Avg_height, fill=team)) +
  geom_bar(stat='identity') +
  coord_cartesian(ylim = c(66,74)) +
  geom_hline(aes(yintercept=ncaa_avg_height,linetype = "NCAA Average Height")) +
  geom_text(aes(label = Avg_height), vjust = 1.5,
              position = position_dodge(width = 0.9))+
  xlab("Team") +
  ylab("Average Height (in)")
```



Extending the Examples (by Kory Martin):

For this we will extend upon the examples presented, by looking at the number of international students that are attending the universities and gleaning some meaningful insights based on this.

We begin by using the **mutate** function in dplpyr to create a new column that will hold a value of 'international' if the student is not from the US, and a value of 'domestic' if they are from the US. In addition to the mutate function, we will use the **if_else** function to assign the value to the new column, based on whether or not the value of country_clean is 'USA' or not.

```
(ncaa_wbb_roster_extended <- ncaa_wbb_rosters %>%
  mutate(domestic_international = ifelse(country_clean == 'USA', 'domestic', 'international')))
```

```
## # A tibble: 13,806 x 30
##    ncaa_id team      player_id name         year  hometown homestate high_school
```

```
##      <dbl> <chr>         <dbl> <chr>        <chr> <chr>    <chr>     <chr>
##  1     721 Air Force     11807 Mackenzie Le Fres~ Elk Gro~ Californ~ St. Francis
##  2     721 Air Force     11805 Milahnie Pe~ Fres~ Tampa    Florida   Seffner Ch~
##  3     721 Air Force     11801 Madison Smi~ Soph~ Connell  Washingt~ Connell
##  4     721 Air Force     11795 Taylor Britt Juni~ Columbia South Ca~ Spring Val~
##  5     721 Air Force     11796 Kamri Heath  Seni~ Edmond   Oklahoma  Edmond Nor~
##  6     721 Air Force     11800 Kayla Pilson Juni~ Houston  Texas     Westside
##  7     721 Air Force     11806 Faith Shelt~ Fres~ Stockton Californ~ Lincoln
##  8     721 Air Force     11804 Griffin Gre~ Fres~ Monument Colorado  Lewis-Palm~
##  9     721 Air Force     11803 Parker Brown Fres~ Spokane  Washingt~ Mead
## 10     721 Air Force     11798 Dasha Macmi~ Juni~ Colleyv~ Texas     Grapevine
## # i 13,796 more rows
## # i 22 more variables: previous_school_clean <chr>, height_clean <chr>,
## #   position <chr>, jersey <chr>, url <chr>, season <chr>, team_state <chr>,
## #   conference <chr>, division <chr>, height_ft <dbl>, height_in <dbl>,
## #   total_inches <dbl>, primary_position <chr>, secondary_position <chr>,
## #   position_clean <chr>, year_clean <chr>, redshirt <dbl>, hs_clean <chr>,
## #   hometown_clean <chr>, state_clean <chr>, country_clean <chr>, ...
```

We will then use **select** function combined with the **slice_sample** function to confirm that our new column has the expected values, by selecting a random sample of rows and looking at their country_clean and domestic_international values

```
ncaa_wbb_roster_extended %>% select(country_clean, domestic_international) %>% slice_sample(n=15)
```

```
## # A tibble: 15 x 2
##    country_clean domestic_international
##    <chr>         <chr>
##  1 USA           domestic
##  2 USA           domestic
##  3 USA           domestic
##  4 USA           domestic
##  5 USA           domestic
##  6 USA           domestic
##  7 USA           domestic
##  8 USA           domestic
##  9 USA           domestic
## 10 USA           domestic
## 11 USA           domestic
## 12 USA           domestic
## 13 USA           domestic
## 14 USA           domestic
## 15 USA           domestic
```

Next we will use the **group_by** function to generate a count of the number of domestic vs international students at each college. Furthermore, we will use the **summarize** function to create summary measures for the total number of domestic students, the total number of international students, and the total number of players on the team. Finally, based on these counts, we will then create an additional variable pct_international that will represent the pct of international students on the student rosters.

```
(ncaa_international_domestic <- ncaa_wbb_roster_extended %>%
  group_by(team) %>%
  summarize(num_domestic = sum(ifelse(domestic_international == 'domestic',1,0)),
```

```
            num_international = sum(ifelse(domestic_international == 'international',1,0)),
            num_players = n()) %>%
  mutate(pct_international = num_international/num_players))
```

```
## # A tibble: 938 x 5
##    team             num_domestic num_international num_players pct_international
##    <chr>                   <dbl>             <dbl>       <int>             <dbl>
##  1 A&M-Corpus Chri~           10                 5          15             0.333
##  2 Abilene Christi~           13                 0          13             0
##  3 Academy of Art              6                 2           8             0.25
##  4 Adams St.                  14                 0          14             0
##  5 Adelphi                    14                 0          14             0
##  6 Air Force                  13                 0          13             0
##  7 Akron                      10                 2          12             0.167
##  8 Alabama                    13                 0          13             0
##  9 Alabama A&M                15                 0          15             0
## 10 Alabama State              14                 0          14             0
## # i 928 more rows
```

We will use the **left_join** function to connect our data to the division in which the college belongs to

```
(ncaa_international_domestic <- left_join(ncaa_international_domestic, ncaa_wbb_rosters %>% select(team
```

```
## # A tibble: 938 x 6
##    team    num_domestic num_international num_players pct_international division
##    <chr>          <dbl>             <dbl>       <int>             <dbl> <chr>
##  1 A&M-Co~           10                 5          15             0.333 I
##  2 Abilen~           13                 0          13             0     I
##  3 Academ~            6                 2           8             0.25  II
##  4 Adams ~           14                 0          14             0     II
##  5 Adelphi           14                 0          14             0     II
##  6 Air Fo~           13                 0          13             0     I
##  7 Akron             10                 2          12             0.167 I
##  8 Alabama           13                 0          13             0     I
##  9 Alabam~           15                 0          15             0     I
## 10 Alabam~           14                 0          14             0     I
## # i 928 more rows
```

Finally we will use the **group_by** and **summarize** functions, to determine the average pct of international students per team, across teams in each of the three divisions.

```
ncaa_international_domestic %>%
  group_by(division) %>%
  summarize(num_teams = n(),
            avg_pct_international = mean(pct_international))
```

```
## # A tibble: 3 x 3
##   division num_teams avg_pct_international
##   <chr>        <int>                 <dbl>
## 1 I              360                0.152
## 2 II             245                0.0735
## 3 III            333                0.0110
```

Conclusion:

As you can see Indiana has the tallest average height, followed by UConn, Michigan, South Carolina, Rutgers and Stanford. As a huge women's basketball fan, one thing that stands out to me is that all but two of these teams have been nationally ranked in the top 5 this season, and of the two that have not been ranked in the top 5, Rutgers and Michigan, Michigan has been ranked in the top 20, pretty consistently. Rutgers has been going through a rough patch in the absence of their hall-of-fame coach C. Vivian Stringer and their roster only contains 8 players (their average height may be inflated by the lack of players on the roster). However, my first thought after looking at this would be that it seems that height plays some component in basketball (stating the obvious).