Project 3

Andrew Gillock apg2255

This is the dataset used in this project:

```
data <- readr::read_csv(
  'https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2022/2022-03-29/sports.csv
  )
head(data)
```

```
## # A tibble: 6 x 28
##    year unitid institution_name city_txt state_cd zip_text classification_~
##   <dbl>  <dbl> <chr>            <chr>    <chr>       <dbl>            <dbl>
## 1  2015 100654 Alabama A & M U~ Normal   AL          35762                2
## 2  2015 100654 Alabama A & M U~ Normal   AL          35762                2
## 3  2015 100654 Alabama A & M U~ Normal   AL          35762                2
## 4  2015 100654 Alabama A & M U~ Normal   AL          35762                2
## 5  2015 100654 Alabama A & M U~ Normal   AL          35762                2
## 6  2015 100654 Alabama A & M U~ Normal   AL          35762                2
## # ... with 21 more variables: classification_name <chr>,
## #   classification_other <chr>, ef_male_count <dbl>, ef_female_count <dbl>,
## #   ef_total_count <dbl>, sector_cd <dbl>, sector_name <chr>, sportscode <dbl>,
## #   partic_men <dbl>, partic_women <dbl>, partic_coed_men <dbl>,
## #   partic_coed_women <dbl>, sum_partic_men <dbl>, sum_partic_women <dbl>,
## #   rev_men <dbl>, rev_women <dbl>, total_rev_menwomen <dbl>, exp_men <dbl>,
## #   exp_women <dbl>, total_exp_menwomen <dbl>, sports <chr>
```

Link to the dataset: https://github.com/rfordatascience/tidytuesday/blob/master/data/2022/2022-03-29/readme.md

Part 1

Question: Looking at the top 10 most popular collegiate sports, how does the total revenue obtained by both male and female participants differ between sports?

Introduction: This dataset was obtained from the tidytuesday repository. It contains information regarding equality in collegiate sports from 2015-2019. We are interested in answering how total revenue obtained differs between the top 10 most popular sports. The *total_rev_menwomen* variable contains the total revenue earned by both men and women, and the *sports* variable contains the name of each collegiate sport. Using these variables, we can answer the question at hand.

Approach: In order to determine which sports are most popular among college athletes, we first need to figure out the frequency of each reported sport. We can assume that sports that occur more frequently than others in the dataset are considered to be the more popular. The "All Track Combined" category from the *sports* variable was removed since it is representative of multiple sports. Using *fct_infreq()*, we can determine which sports are most commonly reported within the dataset, then we can use *fct_lump_n()* to select the top 10. Next, the revenue percentage from each sport will be added to the existing dataset using *mutate()*. A pie chart is ideal for visualizing proportions, so next we will create relevant measurements for the chart and use *geom_arc_bar()* to create it. Labels containing the relevant percentages of revenue obtained from each sport can then be added to better visualize how each sport contributes to the total revenue obtained. Now that we have visualized the proportions, we can use *geom_boxplot()* to visualize the distribution of revenue for each sport.

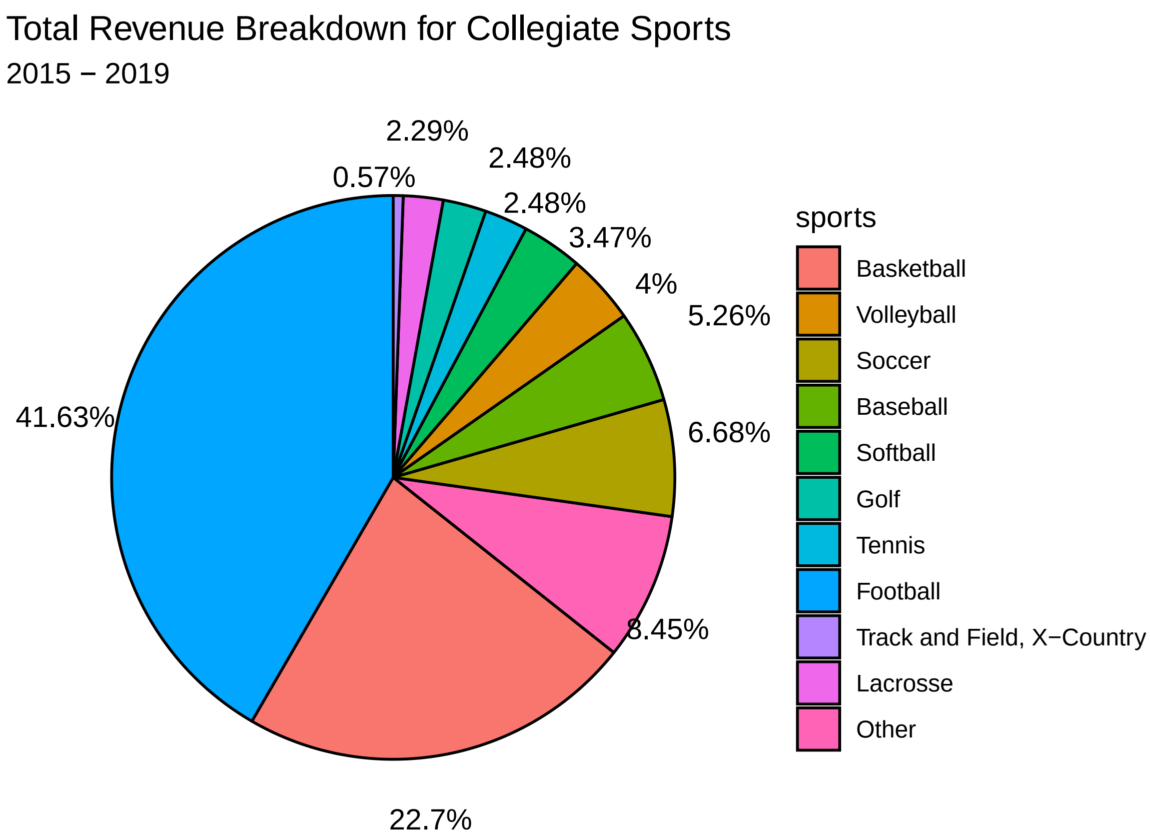
Analysis:

```
top_rev <- data %>%
  filter(sports != "All Track Combined") %>%
  mutate(
  sports = fct_lump_n(fct_infreq(sports), 10),
  other_level = "Other"
  ) %>%
  select(sports, total_rev_menwomen) %>%
  group_by(sports) %>% na.omit() %>%
  summarize(total_rev_menwomen = sum(total_rev_menwomen)) %>%
  mutate(
    perc = (total_rev_menwomen / sum(total_rev_menwomen) * 100)
  )
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

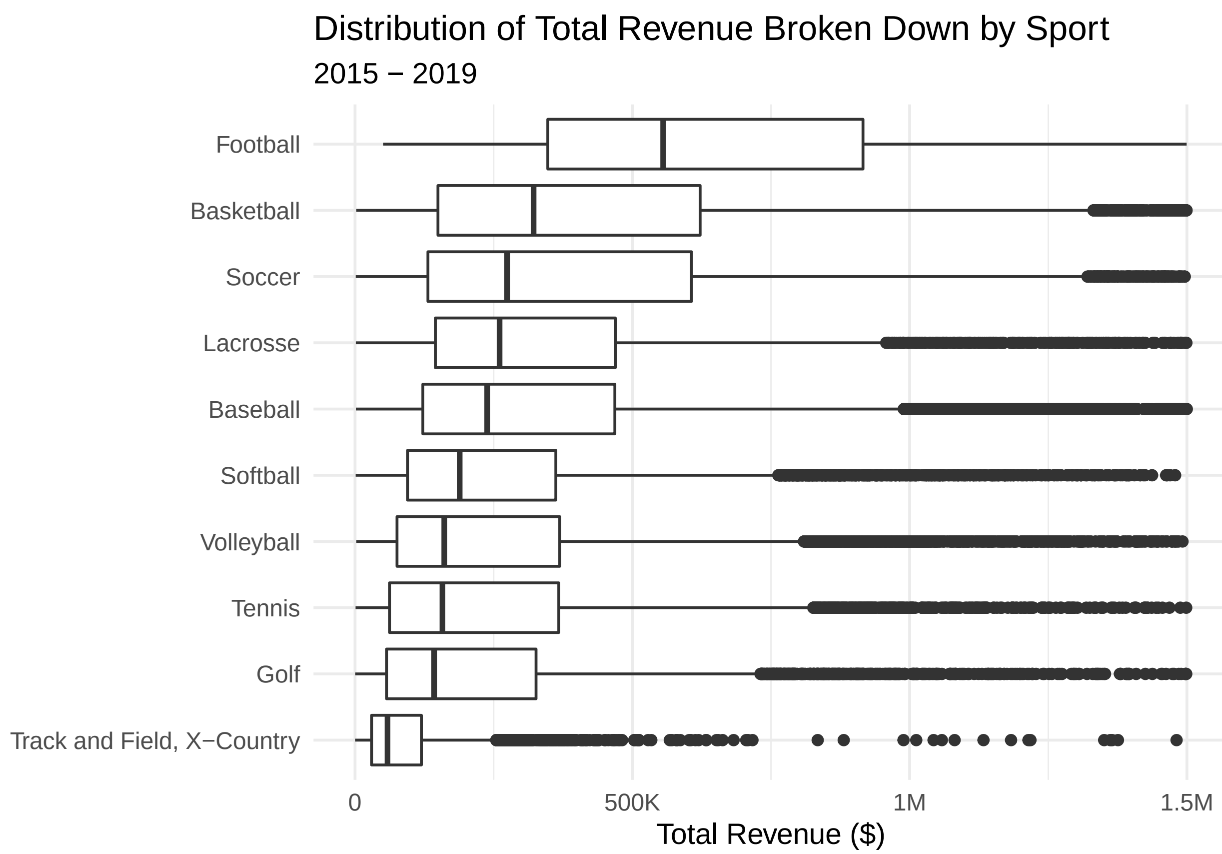
```
pie_data <- top_rev %>% arrange(total_rev_menwomen) %>%
  mutate(
    end_angle = 2*pi*cumsum(perc)/sum(perc),
    start_angle = lag(end_angle, default = 0),
    mid_angle = 0.5*(start_angle + end_angle),
    hjust = ifelse(mid_angle > pi, 1, 0),
    vjust = ifelse(mid_angle < pi/2 | mid_angle > 3* pi/2, 0, 1)
  )

ggplot(pie_data) +
  aes(
    x0 = 0, y0 = 0, # position of pie center
    r0 = 0, r = 2,  # inner and outer radius
    amount = total_rev_menwomen, # size of pie slices
    fill = sports
  ) +
  geom_arc_bar(stat = "pie") +
  geom_text_repel( # place amounts inside the pie
    aes(
      x = 2.3 * sin(mid_angle),
      y = 2.3 * cos(mid_angle),
      label = paste0(round(perc, 2),"%")
    )
  ) +
  coord_fixed(
    xlim = c(-2.5, 2.5), ylim = c(-2.5, 2.5)
  ) + theme_void() +
  ggtitle("Total Revenue Breakdown for Collegiate Sports", "2015 - 2019")
```




```r
#create boxplots for the top 10 sports total revenue gained
top_sports <- data %>%
  group_by(sports) %>% count() %>%
  arrange(desc(n)) %>%
  filter(sports != "All Track Combined")
top10 <- top_sports$sports[1:10] #extract most popular sports from data

data %>% select(sports, total_rev_menwomen) %>%
  na.omit() %>% #remove NAs
  filter(
    sports %in% top10,
    total_rev_menwomen < 1500000 #filter to remove outlying datapoints
    ) %>%
  mutate(
  sports = fct_reorder(sports, total_rev_menwomen, median) #order sports by median
) %>%
  ggplot(aes(x = total_rev_menwomen, y = sports)) + geom_boxplot() +
  scale_x_continuous(name = "Total Revenue ($)",
                     labels = c("0", "500K", "1M", "1.5M")) +
  scale_y_discrete(name = NULL) +
  theme_minimal() +
  ggtitle("Distribution of Total Revenue Broken Down by Sport", "2015 - 2019")
```



Discussion: The pie chart depicts each sport's contribution to the total revenue obtained within the entirety of the dataset. We can see that football and basketball raise 41.63% and 22.7% of the total revenue, respectively. Additionally, the distribution of total revenue suggests a similar result, where football and basketball have the highest median revenue raised. I would assume that this trend exists because these are oftentimes the most popular sports among colleges, which explains how they raise the most money.

Part 2

Question: How well can total revenue earned by the 5 most popular sports be predicted by total student body of the college? To answer this question, create a table of summary statistics for each of the 5 most popular sports. Next, create a plot of the relationship between student body and total revenue earned for each sport. Using the information from the summary table, label the plot with the strength of each model.

Introduction: Using the same dataset as before, we are now interested in predicting the total revenue obtained from each sport depending on the total number of students at the university. We will again use the *sports* and *total_rev_menwomen* variables to answer this question, as well as *ef_total_count*, which contains the total student body for binary male/females.

Approach: To answer this question, we must first mean-center the *ef_total_count* variable to obtain more accurate results. Next, a summary of linear models can be created by nesting the data by the *sports* variable and the *map()* function, which is necessary to analyze models for each respective sport. From there, the top 5 sports can be selected using the *slice()* function. Once the summary table has been generated, we can use the resulting Rˆ2 values to create our label data. Finally, we can plot the lm line for each sport by faceting the plot.

Analysis:

```r
data2 <- data %>% select(sports, total_rev_menwomen, ef_total_count) %>% na.omit()
sum(is.na(data2)) #confirm no NAs
```

```
## [1] 0
```

```r
#scale student count for better analysis
data2$ef_total_count <- scale(
  data2$ef_total_count, scale = FALSE
  )

#create summary
lm_summary <- data2 %>%
  nest(data = -sports) %>%
  mutate(
    fit = map(data, ~lm(total_rev_menwomen ~ ef_total_count, data = .x)),
    glance_out = map(fit, glance)
  ) %>%
  select(sports, glance_out) %>%
  unnest(cols = glance_out) %>%
  arrange(desc(r.squared)) %>% #arrange by largest correlation
  slice_head(n = 5)
lm_summary
```

```
## # A tibble: 5 x 13
##   sports r.squared adj.r.squared  sigma statistic  p.value    df  logLik    AIC
##   <chr>      <dbl>         <dbl>  <dbl>     <dbl>    <dbl> <dbl>   <dbl>  <dbl>
## 1 Footb~     0.513         0.513 1.16e7    4667.  0.            1 -7.85e4 1.57e5
## 2 Baske~     0.393         0.393 2.70e6    6447.  0.            1 -1.61e5 3.23e5
## 3 Squash     0.338         0.334 2.07e5      81.6 5.11e-16      1 -2.21e3 4.43e3
## 4 Volle~     0.261         0.261 3.16e5    3143.  0.            1 -1.25e5 2.50e5
## 5 Baseb~     0.248         0.248 4.80e5    2753.  0.            1 -1.21e5 2.42e5
## # ... with 4 more variables: BIC <dbl>, deviance <dbl>, df.residual <int>,
## #   nobs <int>
```

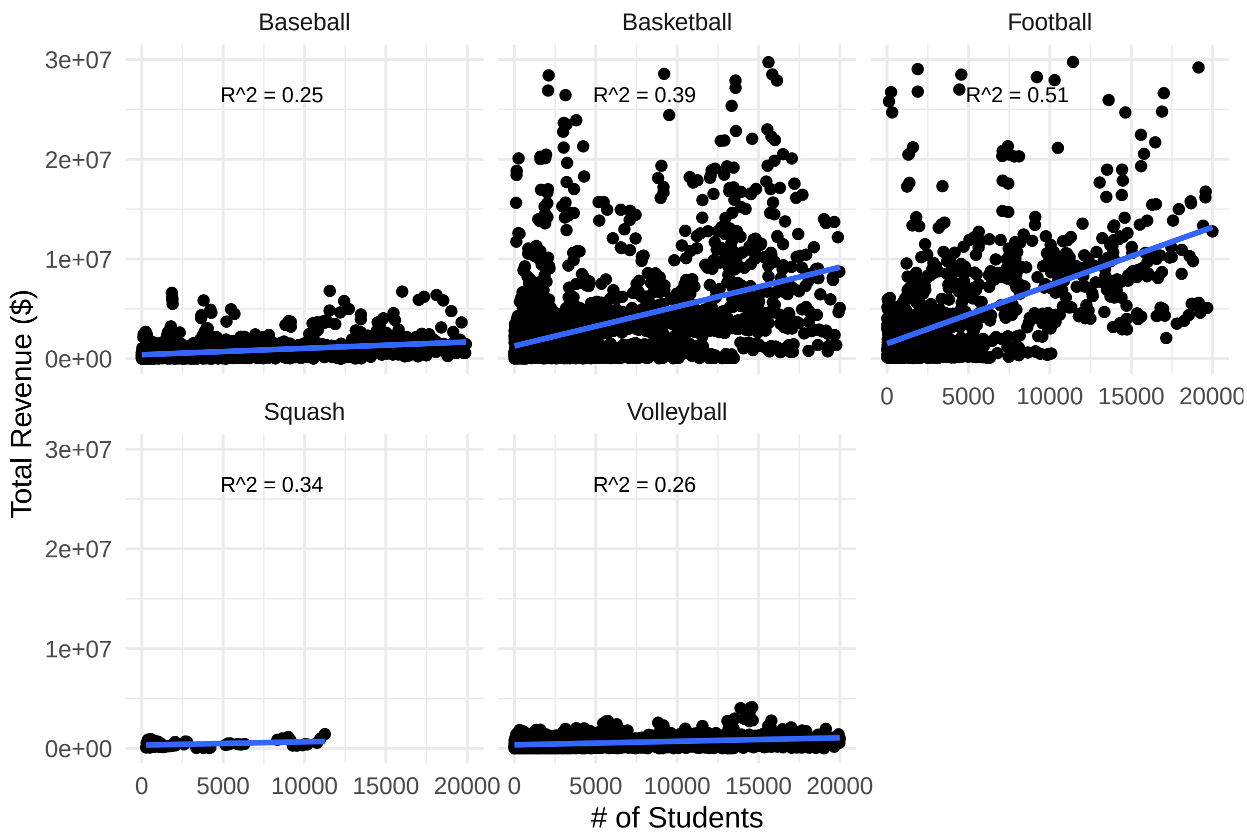
```r
#plot the relationships for the top 5 sports
#create label data
label_data <- lm_summary %>%
  mutate(
    rsqr = signif(r.squared, 2),  # round to 2 significant digits
    label = glue("R^2 = {rsqr}"),
    total_rev_menwomen = 26500000, ef_total_count = 8000 # label position in plot
  ) %>%
  select(sports, label, total_rev_menwomen, ef_total_count)
label_data
```

```
## # A tibble: 5 x 4
##   sports     label      total_rev_menwomen ef_total_count
##   <chr>      <glue>                  <dbl>          <dbl>
## 1 Football   R^2 = 0.51           26500000           8000
## 2 Basketball R^2 = 0.39           26500000           8000
## 3 Squash     R^2 = 0.34           26500000           8000
## 4 Volleyball R^2 = 0.26           26500000           8000
## 5 Baseball   R^2 = 0.25           26500000           8000
```

```r
#and plot
data2 %>% filter(sports %in% c("Football", "Basketball", "Squash", "Volleyball", "Baseball")) %>%
```

```r
ggplot(aes(x = ef_total_count, y = total_rev_menwomen)) +
  geom_point() + geom_text(
  data = label_data, aes(label = label),
  size = 8/.pt #8pt font
) + geom_smooth(method = "lm", se = FALSE) + facet_wrap(~sports) +
theme_minimal() + scale_x_continuous(name = "# of Students", limits = c(0, 20000)) +
scale_y_continuous(name = "Total Revenue ($)", limits = c(0, 30000000)) #set coordinates
```

```
## `geom_smooth()` using formula 'y ~ x'
## Warning: Removed 24216 rows containing non-finite values (stat_smooth).
## Warning: Removed 24216 rows containing missing values (geom_point).
```



Discussion: After analyzing the summary table, we can see that the strongest correlation between student body and total revenue earned is for football and basketball. This is because these sports have the largest R^2 value, which suggests that the model explains 51% of the variation for the football, and 39% for basketball. The lowest correlation occurs for squash, volleyball, and baseball. These conclusions are seen more clearly in the graphs for each sport, where those with high correlation have a steeper line and a higher R^2 value. I would assume that the correlation is stronger in football and basketball because these are the most popular sports within the dataset. Since they are the most popular, it makes sense that they would raise more money than sports such as squash, volleyball, and baseball.