Data Analytics Skills Challenge - Penn Interactive

Stephanie Orgill

4/12/2022

Introduction

Welcome to the Penn Interactive Data Analytics skills challenge. Please keep in mind as you proceed there is no single right answer. Please limit yourself to 1-3 hours on this skills challenge. If you feel you need more time, please reach out because you may be trying to solve more than necessary. This is not Penn Interactive's data but is a great representation of the type of work we do. If this public dataset is no longer available, please let us know.

We want to evaluate two main criteria:

- Can you handle the technical and visualization skills of this role?
- Can you create some interesting insights as a part of your analysis?

There are two steps.

1. Access and query some data from Google's public data set on NCAA basketball.
2. Import the data into a visualization tool of your preference and create some supporting visualizations for your insights.

Data

Here is a link to the NCAA basketball dataset (https://console.cloud.google.com/marketplace/product/ncaa-bb-public/ncaa-basketball).

This dataset contains data about NCAA Basketball games, teams, and players. Game data covers play-by-play and box scores back to 2009, as well as final scores back to 1996. Additional data about wins and losses goes back to the 1894-5 season in some teams' cases.

Pulling data with the below SQL code:

```
SELECT
  game_id, season,venue_name, venue_city, venue_state, attendance, conference_game, tournament,
tournament_type, round, game_no, away_name, away_alias, away_market, away_division_alias, home_n
ame, home_alias, home_market, home_division_alias, elapsed_time_sec, timestamp, team_name, team_
market, team_alias, team_conf_alias, event_description, event_type, shot_type, points_scored
FROM
  `bigquery-public-data.ncaa_basketball.mbb_pbp_sr`
WHERE
  home_division_alias = "D1"
  AND away_division_alias = "D1"
```

```
# load pulled data
ncaa_dataset <- read_csv("C:/Users/orgil/OneDrive/Desktop/Work/Work_challenges/ncaa_dataset.csv"
)
```

```
## Rows: 4119733 Columns: 29
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (24): game_id, venue_name, venue_city, venue_state, tournament, tourname...
## dbl  (4): season, attendance, elapsed_time_sec, points_scored
## lgl  (1): conference_game
##
## 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.
```

Question

For Division 1 basketball teams, how much of an advantage is home court?

Remember, you have freedom to explore the data and create some insights! Some tips:

- Not all home teams play at their home venue (arena / stadium).
    - Some games are a part of a tournament and played at a neutral site venue.
- What does home court advantage mean? You tell us! (Some examples listed below.)
    - Is the number of wins a team has at the home venue vs the away & neutral venues much different?
    - Do teams have more second half comebacks when playing at their home venue? (Home team is losing at the end of the first period/half but wins the game.)
    - Does the team (score more points, have a better field goal percentage, make more free throws, have less turnovers, have less fouls, etc.) when they are at home vs when they are away or neutral venues?
    - Which teams have the best home court advantage? Which teams have the worst? Are there any reasons why this could be the case?
    - Is there seasonality to home court advantage? (Compare the season win rates at home year over year.)
    - Does home court advantage have any impact on ranked teams? (Whether the ranked team is the home team or the away team?)

```
# get unique game info count
length(unique(ncaa_dataset$game_id))
```

```
## [1] 11059
```

```
# replace NA points with 0
ncaa_dataset$points_scored[is.na(ncaa_dataset$points_scored)] <- 0

# get sum of points by game
point_totals_by_game <- ncaa_dataset %>%
  group_by(game_id, home_name, away_name, team_name) %>%
  summarize(total_score = sum(points_scored), .groups = 'drop') %>%
  na.omit() # omit any empty sum rows
# 33,104 -> 22,046
point_totals_by_game
```

```
## # A tibble: 22,046 x 5
##    game_id                              home_name away_name team_name total_score
##    <chr>                                <chr>     <chr>     <chr>           <dbl>
##  1 000872e5-f02a-4b64-ac73-b6b1a7ad10~ Gators    Bulldogs  Bulldogs           72
##  2 000872e5-f02a-4b64-ac73-b6b1a7ad10~ Gators    Bulldogs  Gators             69
##  3 000918c3-b8bf-472a-9a12-94a1984700~ Tigers    Cardinals Cardinals          66
##  4 000918c3-b8bf-472a-9a12-94a1984700~ Tigers    Cardinals Tigers             72
##  5 000b3698-3ce9-44bd-954b-65703a2c6e~ Bulldogs  Razorbac~ Bulldogs           66
##  6 000b3698-3ce9-44bd-954b-65703a2c6e~ Bulldogs  Razorbac~ Razorbac~          61
##  7 00114857-a305-499f-95d2-2b17622666~ Patriots  Rams      Patriots           82
##  8 00114857-a305-499f-95d2-2b17622666~ Patriots  Rams      Rams               71
##  9 0012eab2-237c-4df1-846b-3643053086~ Tigers    Jaguars   Jaguars            66
## 10 0012eab2-237c-4df1-846b-3643053086~ Tigers    Jaguars   Tigers             78
## # ... with 22,036 more rows
```

```
# get summary information about point totals
summary(point_totals_by_game$total_score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   62.00   70.00   70.94   79.00  384.00
```

```
# median: 70
# mean: 70.94

# add column indicating if the point scorer was the away or home team
point_totals_2 <- point_totals_by_game %>%
  mutate(point_scorer = if_else(team_name == home_name, "home", "away")
  )
# remove team_name rows
point_totals_clean <- subset(point_totals_2, team_name != "team_name")
# convert to dataframe object

# get summary info by home or away status
tapply(point_totals_clean$total_score, point_totals_clean$point_scorer, summary)
```

```
## $away
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   25.00   59.00   67.00   67.17   75.00  304.00
##
## $home
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    26.0    66.0    74.0    74.7    83.0   384.0
```

```
# away mean: 67.17
# home mean: 74.7
# ~ difference: 7.53
# away median: 67
# home median: 74
# ~ difference: 7

# get list of the top teams with highest average point totals
tapply(point_totals_clean$total_score, point_totals_clean$team_name, mean) %>%
  sort(,decreasing = TRUE) %>%
  head(5)
```

```
## Black Knights     Tar Heels        Bruins   Blue Devils      Jayhawks
##      81.60000      80.59677      79.49223      79.28426      79.01639
```

```
# highest average score: 81.60 Black Knights

# get list of the bottom teams with the lowest average point
tapply(point_totals_clean$total_score, point_totals_clean$team_name, mean) %>%
  sort(,decreasing = TRUE) %>%
  tail(5)
```

```
##       Braves   Midshipmen    Red Foxes    Blue Hose Golden Lions
##     58.30233     57.81250     57.20000     54.44444     52.87500
```

```
# lowest average score: 52.87 Golden Lions

# list of top and bottom average scoring teams
top_bottom_list <- c("Black Knights", "Tar Heels", "Bruins", "Blue Devils", "Jayhawks", "Braves"
, "Midshipmen", "Red Foxes", "Blue Hose", "Golden Lions")

# get set of games from top and bottom teams
top_bottom_10 <- point_totals_clean[point_totals_clean$team_name %in% top_bottom_list,]
top_bottom_10
```

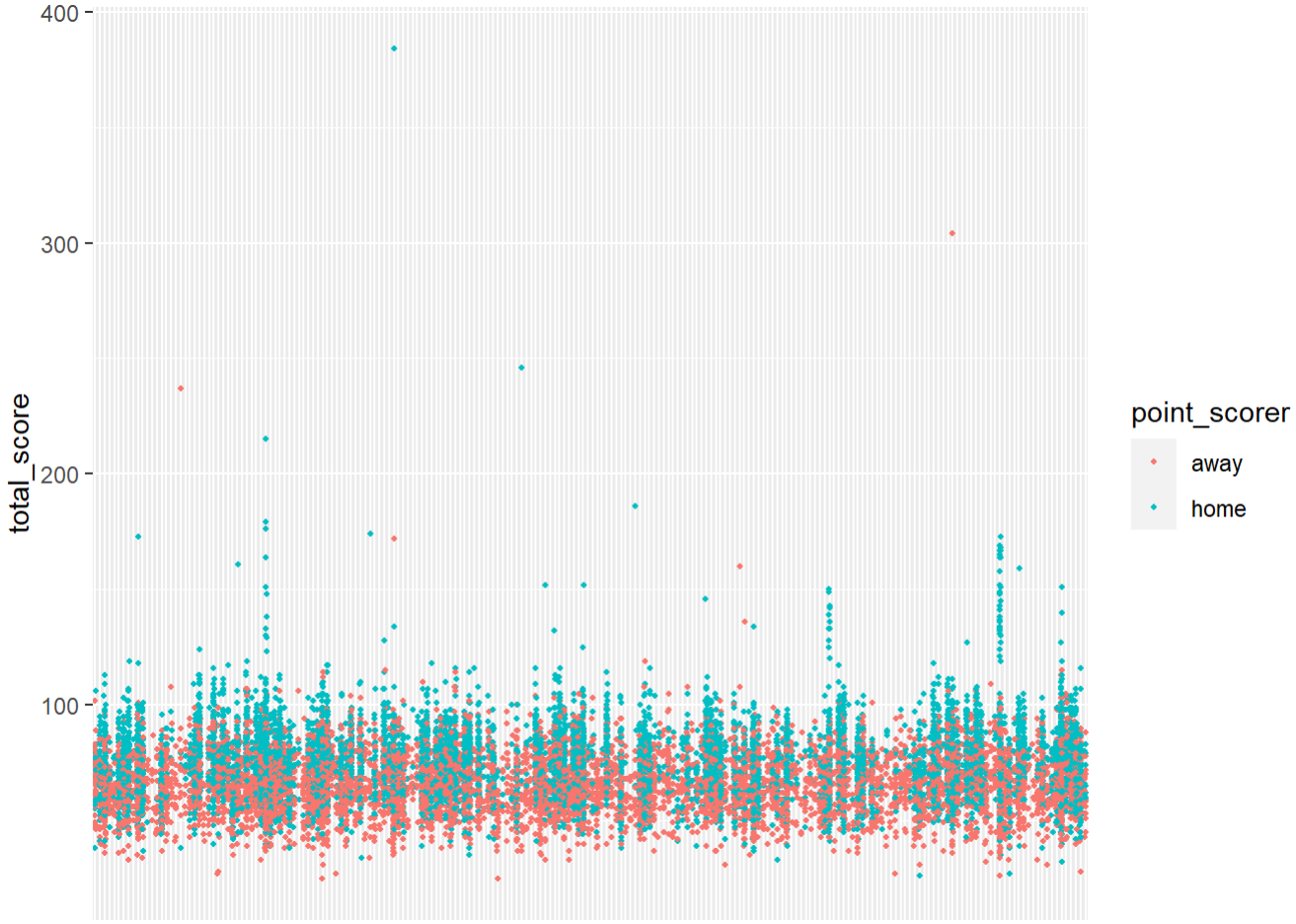
```
## # A tibble: 919 x 6
##    game_id                home_name away_name team_name total_score point_scorer
##    <chr>                  <chr>     <chr>     <chr>           <dbl> <chr>
##  1 005c50e1-c246-474e-98~ Tar Heels Razorbac~ Tar Heels          72 home
##  2 009a851f-1ca7-4ba7-91~ Blue Dev~ Hokies    Blue Dev~          66 home
##  3 00bb32d8-67a2-4c8a-a3~ Tar Heels Blue Dev~ Blue Dev~          93 away
##  4 00bb32d8-67a2-4c8a-a3~ Tar Heels Blue Dev~ Tar Heels          83 home
##  5 00c021e8-e1ac-4081-8e~ Bruins    Huskies   Bruins             98 home
##  6 011ef02a-6c40-4833-b3~ Tar Heels Panthers  Tar Heels          85 home
##  7 014c0b9d-79ee-4602-8a~ Jayhawks  Wildcats  Jayhawks           86 home
##  8 01ece451-d68c-4918-b7~ Braves    Shockers  Braves             58 home
##  9 02b01645-c31d-4ec8-8d~ Demon De~ Blue Dev~ Blue Dev~          91 away
## 10 02b5ae5c-b7c0-4092-8e~ Bruins    49ers     Bruins            114 home
## # ... with 909 more rows
```

Visualization

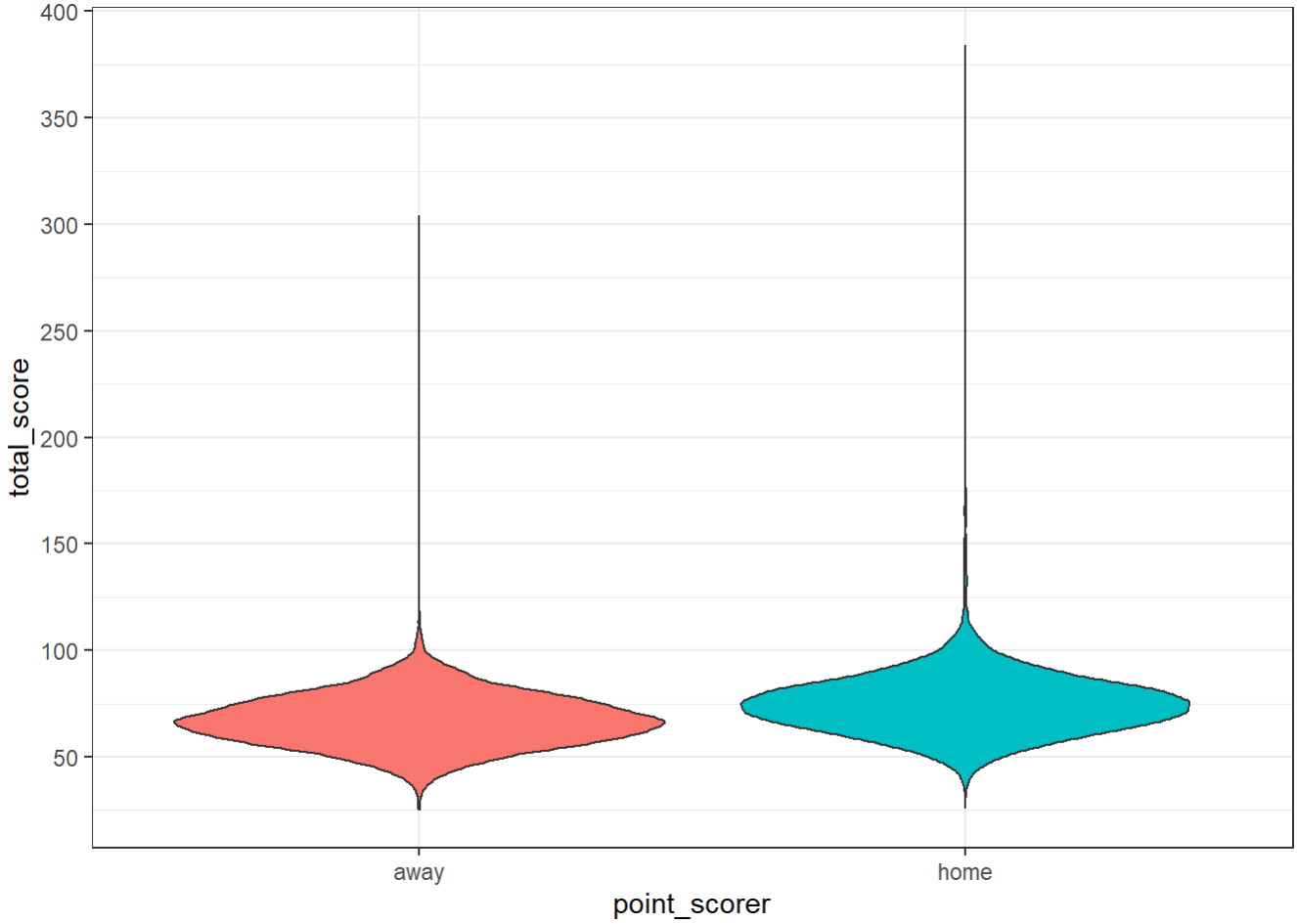
Now show off your visualization skills! Take the data you think is valuable from Google's public NCAA basketball dataset and load it into a visualization tool of your choosing. If you want a free option, you can use the Google Data Studio within the same free account you just created on GCP. After you run your query, click the "Explore Data" & "Explore with Data Studio". Decide which visuals best represent your discovery and insights.

```
# let's visualize the points of home and away teams

# strip chart of teams vs points scored
ggplot(point_totals_clean, aes(x = team_name, y = total_score)) +
  geom_point(size = .75,
             position = position_jitter(
               width = .15, # jitter in horizontal direction
               height = 0 # no vertical jitter
             ),
             aes(color = point_scorer)
  ) +
  theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```

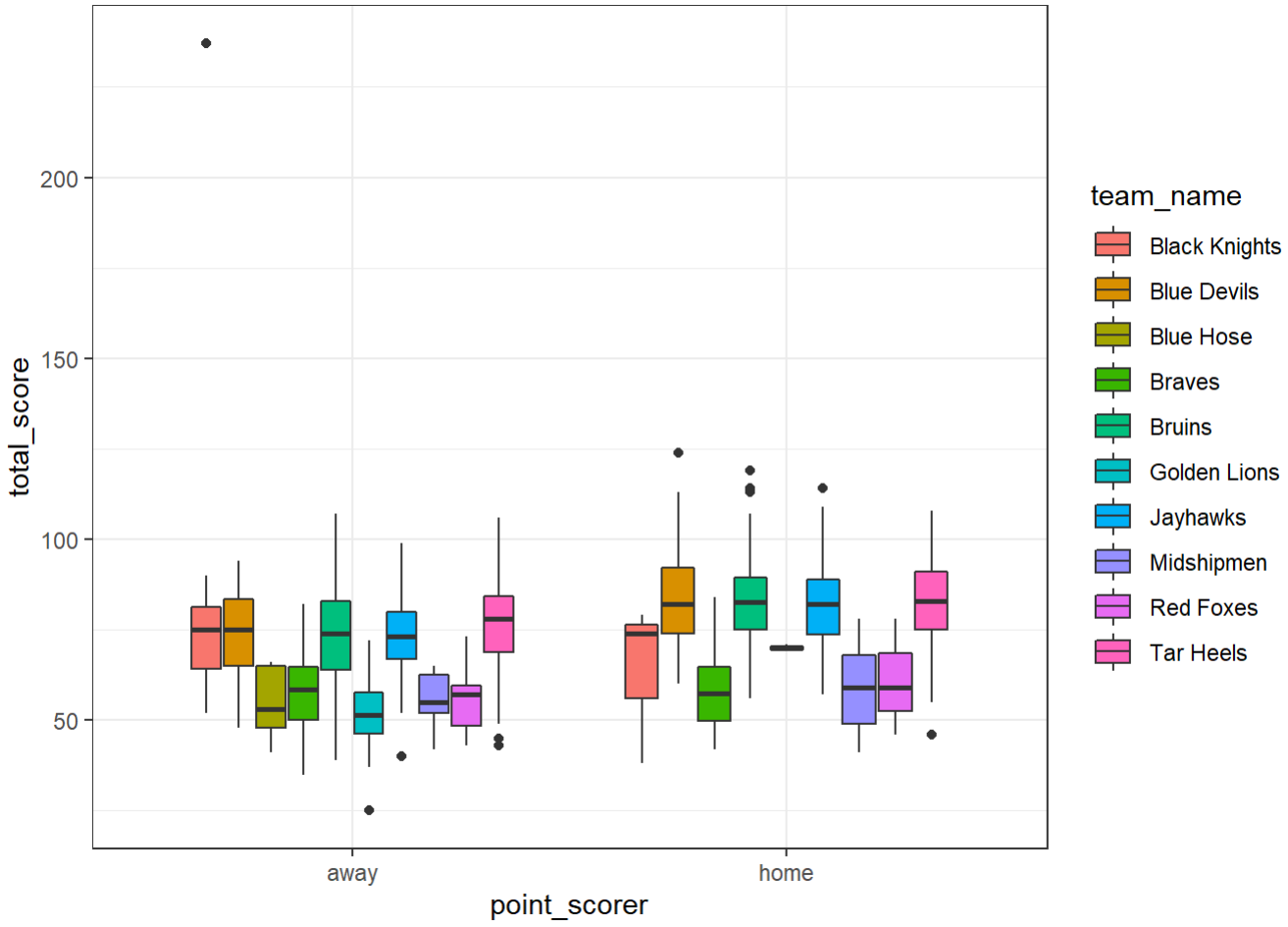



```
# violin plot: home vs away team points
ggplot(point_totals_clean, aes(point_scorer, total_score, fill = point_scorer)) +
  geom_violin() +
  scale_y_continuous(
    breaks = seq(0, 400, 50),
    labels = seq(0, 400, 50)
  ) +
  guides(fill = FALSE) + # remove legend
  theme_bw()
```




```
# on average, home teams tend to score higher than away teams.

# box plot: top and bottom teams - home vs away
ggplot(top_bottom_10, aes(point_scorer, total_score, fill = team_name)) +
  geom_boxplot() +
  scale_y_continuous(
    breaks = seq(0, 200, 50),
    labels = seq(0, 200, 50)
  ) +
  theme_bw()
```



Conclusion

I wanted to focus exclusively on the difference in points for teams playing at home or away. In an effort to do this, I cleaned up the data, calculated the total points for each team for each game. On average, teams playing at their home stadium score 74.7 points while those playing at an away stadium score 67.17 points (a difference of 7.53 points). This relationship is easily seen in the strip chart and the violin plot - on average, home games tend to score higher than away games. There are outliers to this trend, but it seems to remain consistent most of the time.

I wanted to take a look at a more narrow group of teams, so I selected the top 5 and bottom 5 average scoring teams from this dataset.

We can see that there is a great disparity in average score between the best and the worst teams - roughly a range of 28 points per game on average. When we look at the boxplots for these ten teams, we can note that sometimes they buck the expected trend - we expect that home games will generally have higher scores than away games, but it is not always true. For example, the Black Knights have a slightly lower average home score than away score while the Jayhawks score higher at home than they do away on average. For some teams (such as the Blue Hose), I could not make a meaningful comparison between their home and away games due to a lack of datapoints. Based off the top 5 and bottom 5 teams, we can draw some interesting insights about their home/away behavior:

Top 5 teams:

- Black Knights (81.6): scores better at away games
- Tar Heels (80.597): scores better at home games
- Bruins (79.49): scores better at home games
- Blue Devils (79.28): scores better at home games

- Jayhawks (79.016): scores better at home games

Bottom 5 teams:

- Braves (58.30): scores better at away games, but roughly equal
- Midshipmen (57.81): scores better at home games
- Red Foxes (57.20): scores better at home games, but roughly equal
- Blue Hose (54.44): inconclusive
- Golden Lions (52.87): scores better at home games (* based off very small sample size)