R vs. Python: Data Science Comparison

By Tanner Bessette

SYE Goals

- Learn Python for data science
- Go through a Python data science textbook and explore Positron IDE
- Compare R and Python functions side by side
- Be able to identify strengths of both languages, and combine them in a project

What is R? What is Python?

- Both R and Python are coding languages
- R is specifically designed for statistics/data analysis
- Python is utilized for a wider variety of things, including software development, web applications, machine learning, and data science

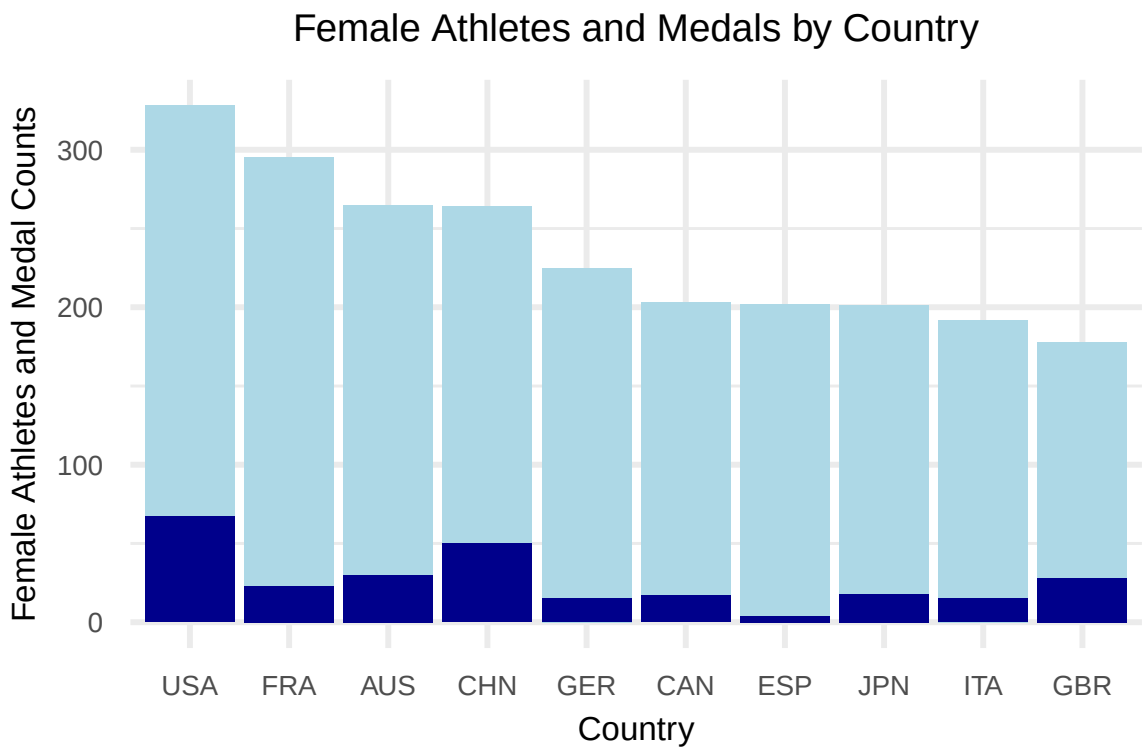
Olympics Project Overview

- Split the semester into two main projects: 2024 Paris Olympics and NBA
- First: Wrangle data and generate plots in R
- Next: replicate everything in Python, following along with a textbook called "Python for Data Science"

Some Olympics Project Plots

Female Athletes Barplot:

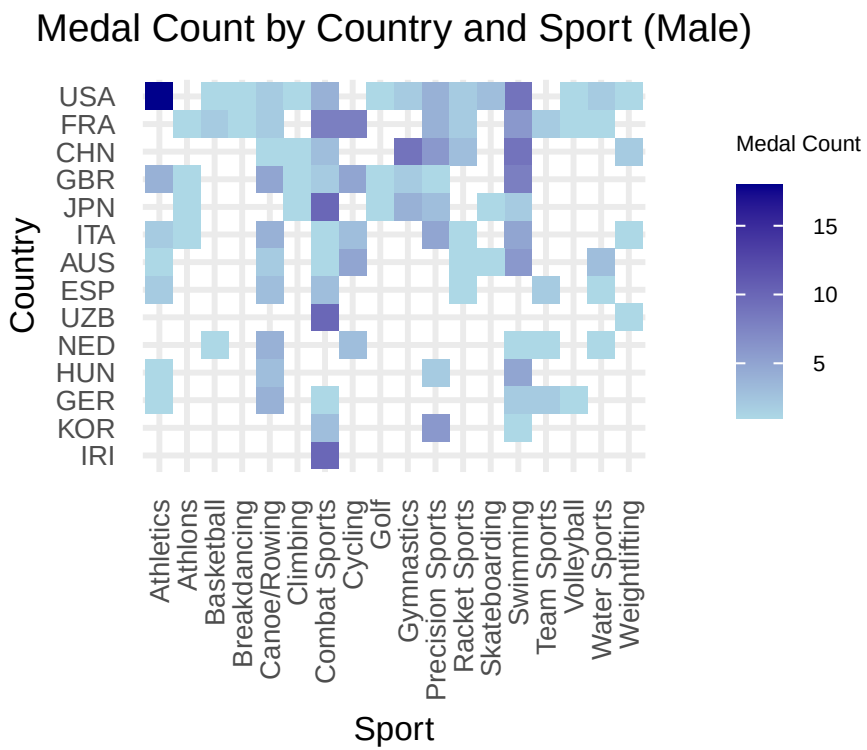
- Light blue represents number of athletes each country sent to the Olympics
- Dark blue represents the number of medals won



Some Olympics Project Plots

Male Medals Heatmap:

- Each heatmap square corresponds to one country and one sport
- The darkness of the color represents the number of medals that country's male athletes won in that sport



R vs. Python Code Comparison

- Here is how I performed the joining of the male athletes and male medals datasets before creating the bar plot using R:

```r
Male_Athletes_Count <-
  left_join(Male_Athletes_Count,
            Male_Total_Medal_Counts,
            by = "country_code")
```

- And here is how I did that in Python:

```python
Male_Athletes_Count_joined = pd.merge(
  Male_Athletes_Count, Male_Total_Medal_Counts,
  on="country_code", how="left")
```

R vs. Python Code Comparison

- Before making the male bar plot, I had to group by country and add up all the athletes for each country.
- Here is how I did it in R:

```r
Male_Athletes_Count <- Male_Athletes_df %>%
  group_by(country_code) %>%
  summarise(athlete_count = n())
```

- And here is how I did it in Python:

```python
Male_Athletes_Count = Male_Athletes_df.groupby(
"country_code").size().reset_index(name="athlete_count")
```

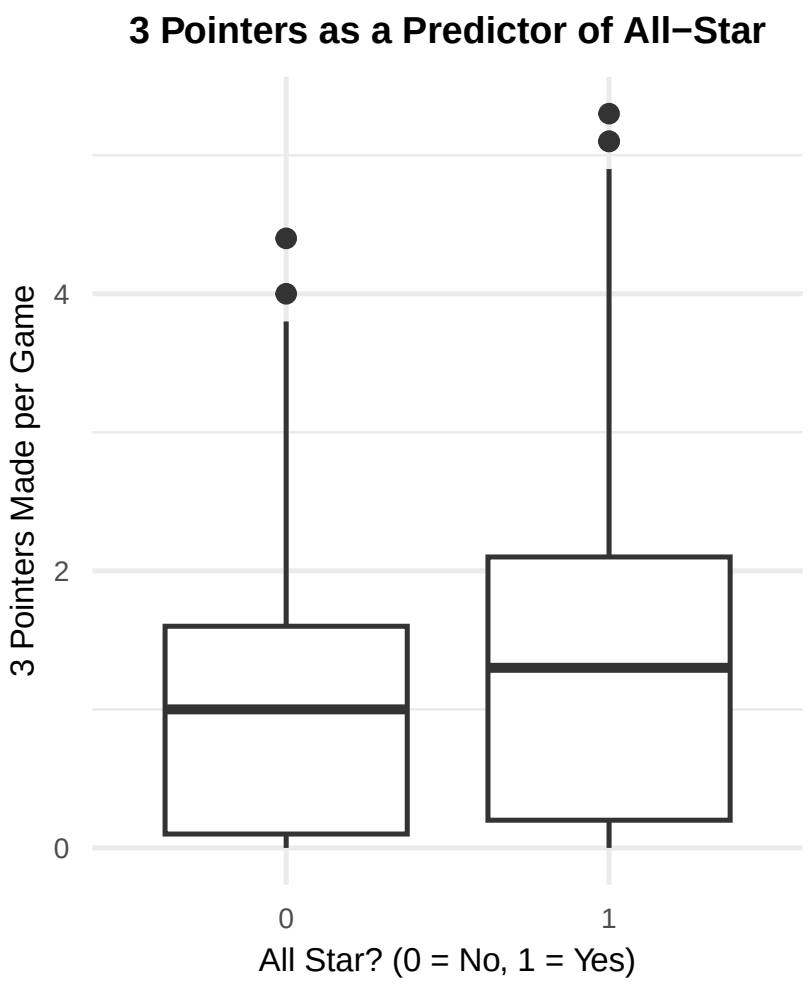
NBA Project

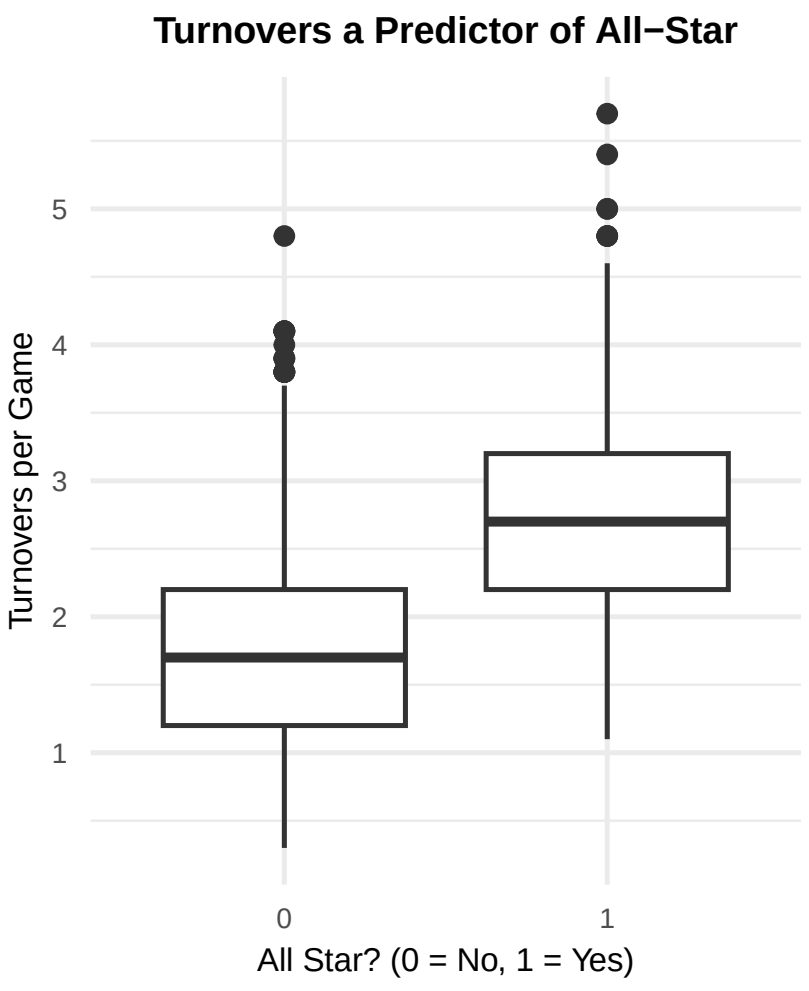
- The second project I worked on was predicting whether NBA players made the all-star game based on their stats
- The dataset included all player statistics each season from 1947 to present
- First part of the project: wrangle and plot in R
- Second part: perform a statistical learning method (KNN) in Python

NBA Project Results

- Interestingly, when making the exploratory plots, one of the variables that looked the most correlated with players making the all-star game was having a high average turnovers
- The variables I used in the KNN model were: points, rebounds, assists, steals, field goals attempted, and turnovers
- Final KNN model accuracy was 88.22%

Turnovers and Three Pointers as Predictors of All-Star





Confusion Matrix

- **Confusion matrix with the results of my KNN model**
- Top left number, 559, gives us the number of non-all-stars that our knn model correctly predicted as non-all-stars
- Bottom right number → number of all-stars that our model correctly predicted as all-stars.
- Bottom left → non-all-stars that our model predicted to actually be all-stars
- Top right → players that our model predicted to not be all-stars but were actually all-stars

```
# Create and display the confusion matrix
conf_matrix = sklearn.metrics.confusion_matrix(
  test_cat, knn_mod)
print(conf_matrix)
```

```
## [[559  14]
##  [ 70  70]]
```

Thoughts on Learning Python for Data Science

- I am surprised that Python is used so much for data science outside of SLU
- R seems more intuitive and organized, especially when it comes to data wrangling
- As an introduction to coding or a machine learning class, I believe Python makes more sense to learn, but I'm glad our department teaches R for DATA234

Conclusion

- In my opinion, R is easier to learn for data science, especially for beginner coders
- Both languages have similar math/statistical capabilities
- For data wranging and plotting, R is great!
- Python may be better (or at least equal) in terms of stat. learning