Advanced methods in applied statistics: Problem set 1

Tomás Fernández Bouvier

February 17, 2021

Total: 9.0 points

1 +2.5 pts

It is generally considered "good practice" to make bins of constant size (and share the same bin edges) when plotting several datasets. You could define your bin range manually, e.g. bin_edges = np.linspace(85,115,15) and pass it directly to plt.hist(samples, bins = bin_edges). This way you could see the contribution of each histogram to the same bin.

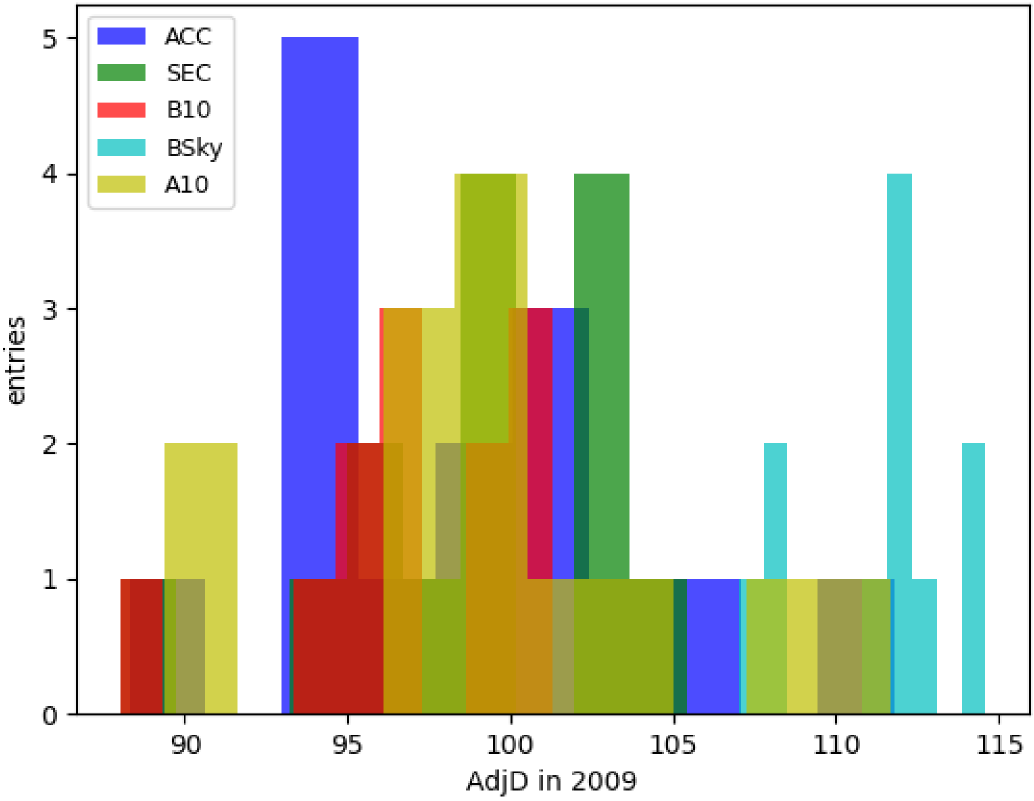

Figure 1: Histogram representation of AdjD for each conference

Further, you did not save the figure as a PDF (I can see it because the image is rasterized (you'd get no decrease in quality from stretching a PDF).

Great job!

2 +3 pts

The results are correct but the report is missing an explanation on how you chose the teams for 2014 and 2009 to compare them (e.g. how you treated the teams that don't appear in both years or change conferences).

The assignment requested a "write-up" which implied some written statements about the steps you took to get to your results… which look awesome! :)

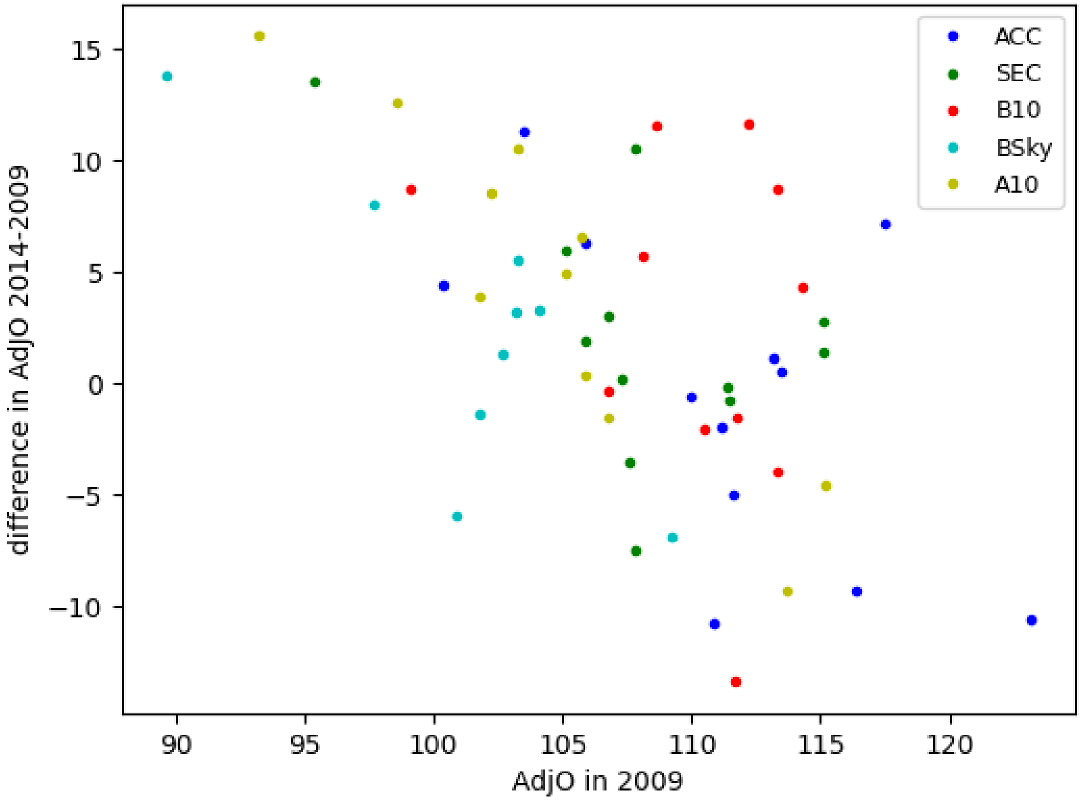

Figure 2: Scatter plot of the imporvement of each team from 2009 to 2014

Conference	means	medians
ACC	-0.625	-0.05
SEC	2.283	1.65
B10	2.673	4.30
BSky	2.322	3.20
A10	4.336	4.90
other	2.595	1.90

✓

Table 1: Table of medians of the difference between 2014 and 2009 in AdjO

3 +3 pts

The numerical results are great, full points here for that! For the future I'd like to recommend constant bin sizes in histograms and some written explanatory statements in the write-ups :)

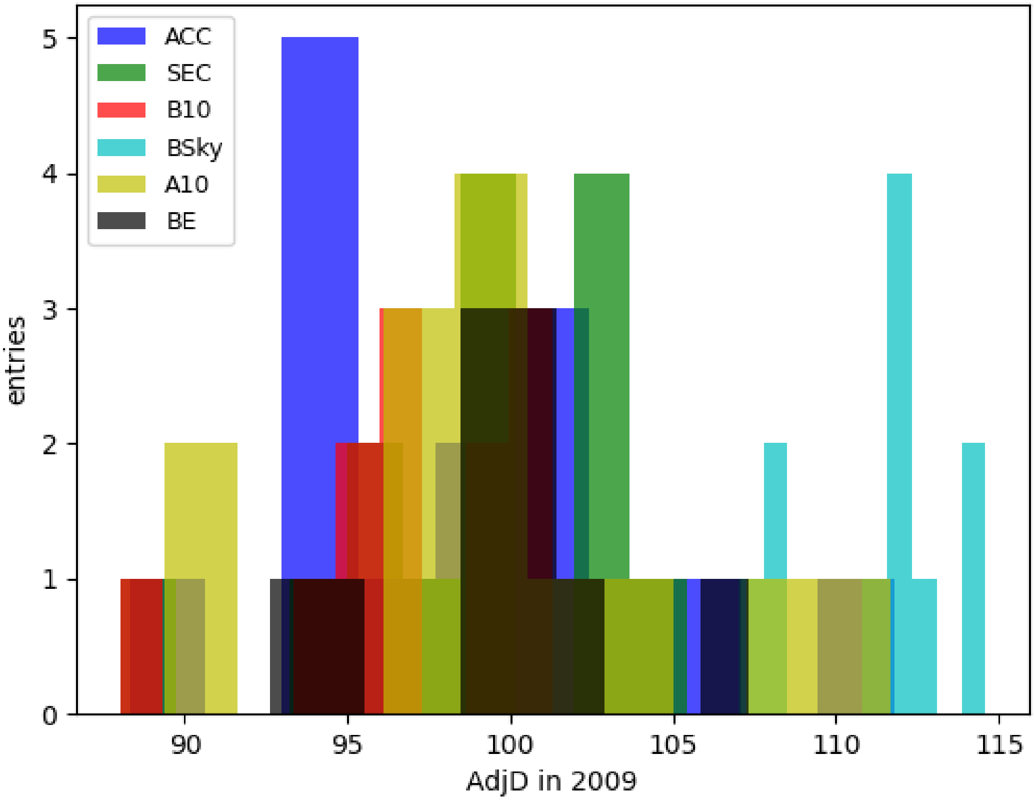

Figure 3: Histogram representation of AdjD for each conference including BE

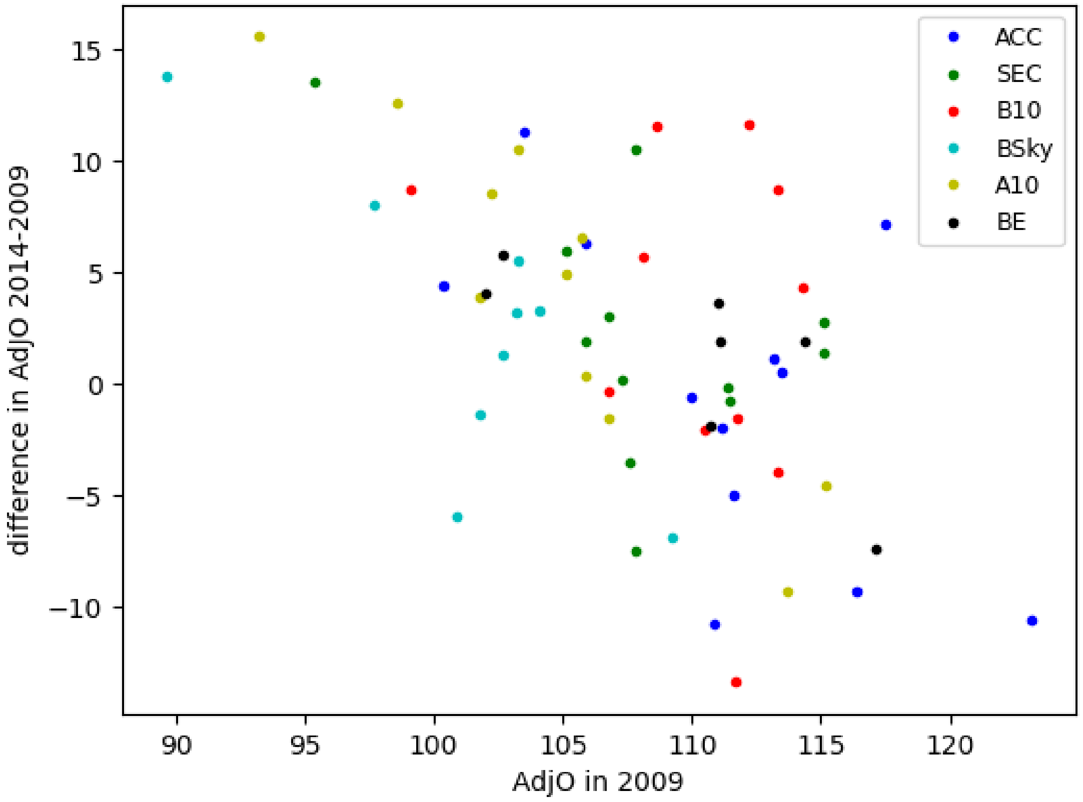

Figure 4: Scatter plot of the imporvement of each team from 2009 to 2014 including BE

Conference	means	medians
ACC	-0.625	-0.05
SEC	2.283	1.65
B10	2.673	4.30
BSky	2.322	3.20
A10	4.336	4.90
BE	1.143	1.90
other	2.624	1.85

Table 2: Table of medians of the difference between 2014 and 2009 in AdjO

4 (EXTRA) +0.5 pts

After parsing and depuring the data in the pdf document we were able to obtain a list of names of authors. The number of unique authors in this list is 3522. After ordering the list aphabetically we also obtained that the mid point is situated around the author "K. DE"

I looked through the code — nice solution! Good thinking with choosing the “set” for uniqueness, and with removing the collaboration names using re < > expressions — please report this in text next time as well. Based on your result, I am guessing that you sorted the list by the first name+last name combination rather than just the last name. It would be great if you included more details about your solutions in the report! (e.g. how you ensured uniqueness, treated the sorting of special characters, whether you sorted in ascending/descending order [I see you did reverse = True; why?], etc so it's easier to benchmark. Good job otherwise!)