Problem Set 1

Emilie Jessen (CPR181)

Advanced Methods in Applied Statistics
14. February 2024

1 Loading the data

The data is loaded with pandas library function: `read_html`, hereby reading the html-files into pandas data frames. The index of the header and the indices for the lines, where the header is repeating itself, are found by inspection. The repeated headers are skipped. To test if the reading of the data was successful, the total number of teams found by pandas each year was compared to the last team number on the web pages. The numbers aligned and the reading of the data was successful.

Exercise 1

Extracting the values for Adjusted Defence 'AdjD' for all the teams participating in the five selected conferences ('ACC', 'SEC', 'B10', 'BSky', 'A10') in 2014. Sorting these values according to conferences results in the five separate histograms for each of the conferences in Fig. 1. The histograms are stacked on top of each other, so the view isn't obscured.

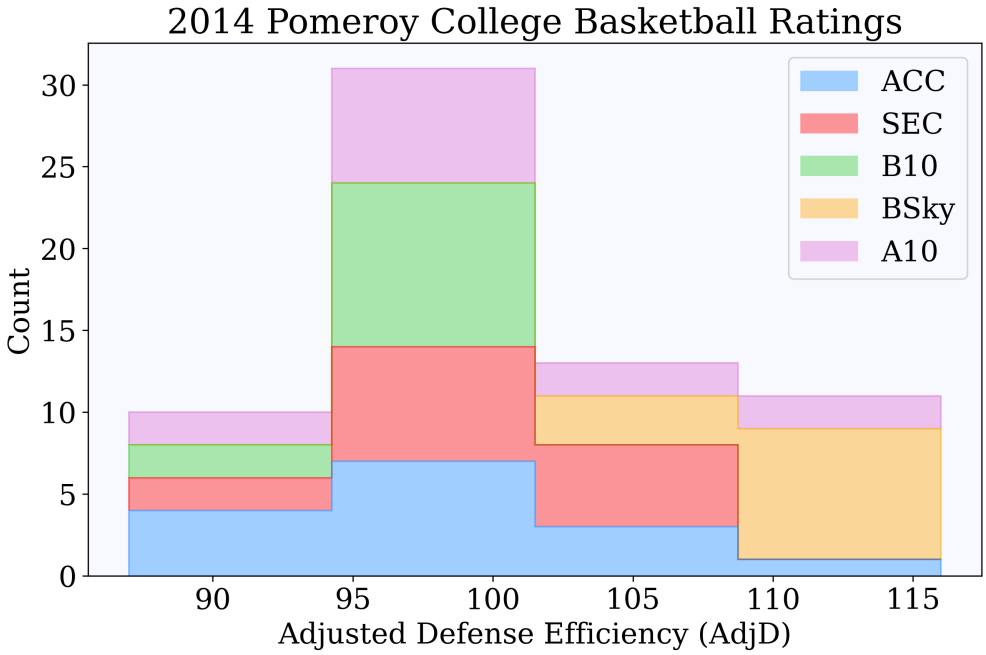

Figure 1: Histograms of Adjusted Defence 'AdjD' in 2014 for all teams participating in the five selected conferences. Each conference has its separate histogram. Histograms are stacked on top of each other.

Exercise 2

To be able to calculate the difference in Adjusted Offence 'AdjO' for the teams in the five conferences in 2009 and 2014, we have to establish:

1. if the teams were participating in both 2009 and 2014.
2. if they were in the same conference both years.

First all characters not in the English alphabet were removed to make the team names for the two years comparable. Then the teams not participating both years were identified. Nine teams were not playing in 2009 and two teams were not in 2014. Therefore the 11 teams were excluded from further analysis.

Second, all teams not in the selected conferences were removed from the analysis. From this 58 teams were identified to participate both years in the selected conferences. Next up was a check if the teams were participating in the same conference both years and if not, they would be excluded from further analysis. Of the 58 teams participating both years in the selected conferences, 55 were in the same conference both years.

For these 55 teams participating both years in the same conference, the Adjusted Offence values were extracted and the difference between 2009 and 2014 were calculated. The differences were sorted into the different conferences to color-code them. The difference in Adjusted Offence as a function of the 2009 value can be seen in Fig. 2.

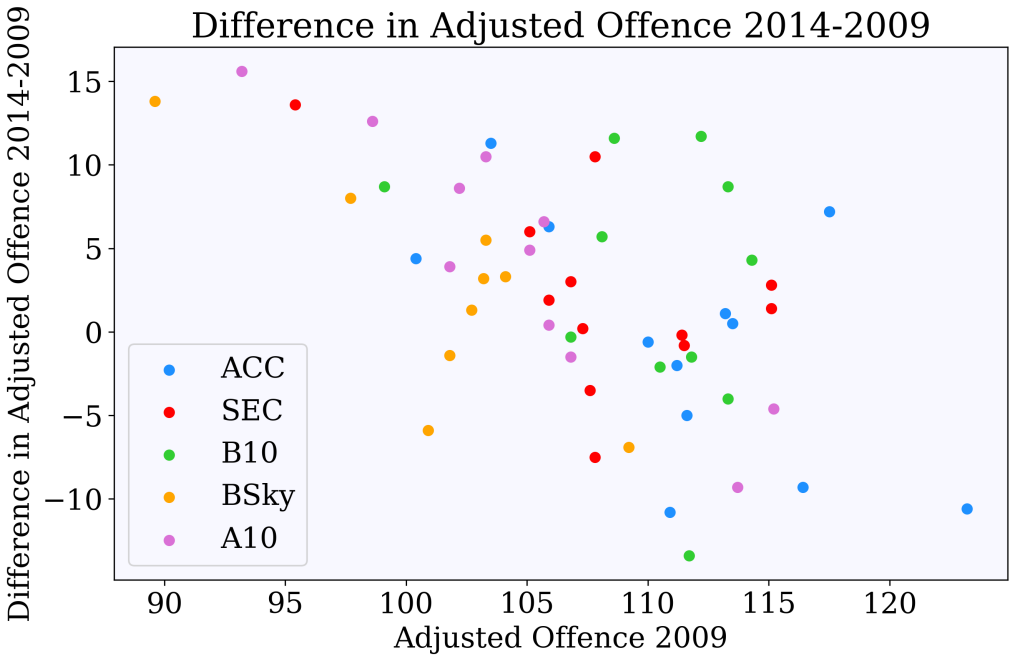

Figure 2: The difference in Adjusted Offence as a function of the 2009 value for each of the five selected conferences.

The mean and median values for the five conferences can be seen in the table below:

Conference	Mean	Median
ACC	-0.625	-0.050
SEC	2.283	1.650
B10	2.673	4.300
BSky	2.322	3.200
A10	4.336	4.900

For the next part of the exercise the conferences not initially selected were identified, and one conference were not part of both 2009 and 2014, so it was excluded. The number of newly selected conferences: 26. The teams not participating both years, not participating in the same conference each year and not in the newly selected conferences were excluded. Of the 274 teams participating both years in the 26 newly selected conferences, 209 were in the same conference both years. For those 209 teams the Adjusted Offence values were extracted and the difference calculated. Sorting them into conferences to calculate the mean and median for each of the 26 conferences. The values can be seen in the table below:

Conference	Mean	Median
AE	-0.800	-0.450
ASun	3.289.	2.400
B12	2.850	0.850
BE	1.143	1.900
BSth	4.200	3.350
BW	1.713	0.800
CAA	4.150	6.900
CUSA	-2.550	-3.650
Horz	2.075	2.300
Ivy	7.137	9.000
MAAC	4.511	4.500
MAC	3.983	3.400
MEAC	2.991	2.200
MVC	2.644	2.600
MWC	1.100	1.150
NEC	2.711	2.400
OVC	0.830	0.600
Pat	7.587	7.700
SB	0.943	0.400
SC	-0.045	0.300
SWAC	2.950	2.700
Slnd	3.067	1.000
Sum	-0.000	1.200
WAC	2.250	2.250
WCC	6.638	6.350
ind	25.100	25.100

Exercise 3

Exercise 1 is repeated with the conference 'BE' included in the initially selected conferences. i.e. six initially selected conferences. This results in the histograms in Fig. 3.

Repeating Exercise 2 with 'BE' included in the initially selected conferences results in the plot in Fig. 4.

The mean and median values for the six initially selected conferences and the remaining 25 conferences align with the values previously reported in Exercise 2.

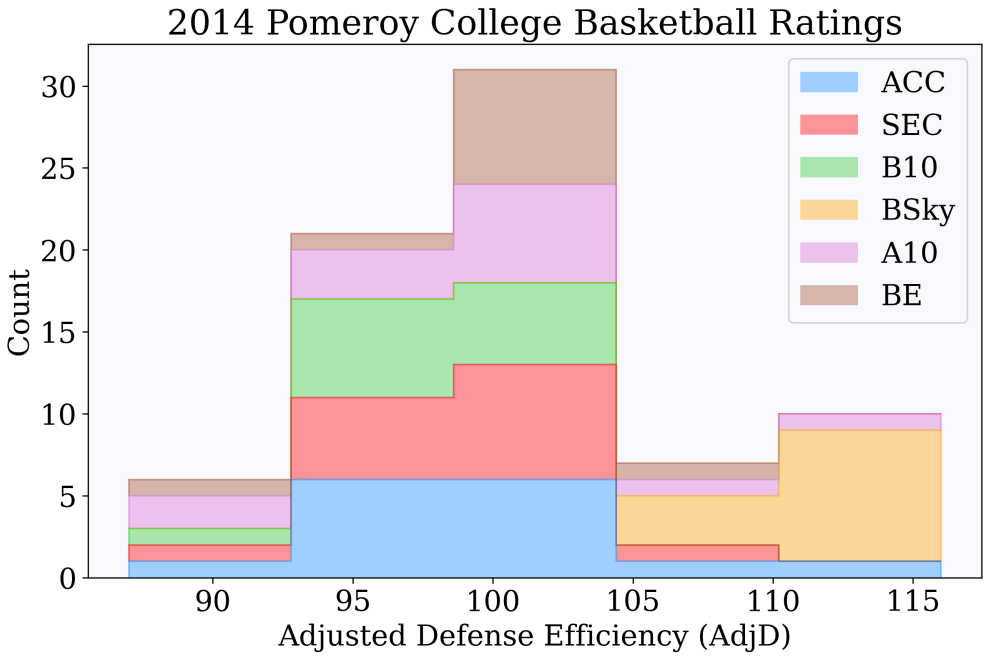

Figure 3: Histograms of Adjusted Defence 'AdjD' in 2014 for all teams participating in the six selected conferences. Each conference has its separate histogram. Histograms are stacked on top of each other.

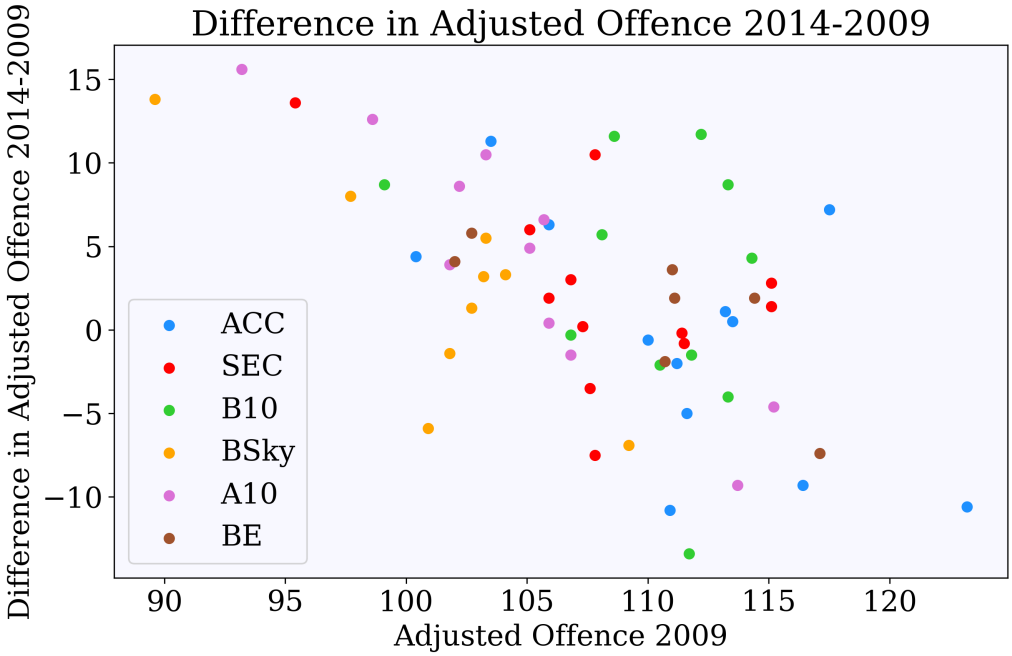

Figure 4: The difference in Adjusted Offence as a function of the 2009 value for each of the five selected conferences.