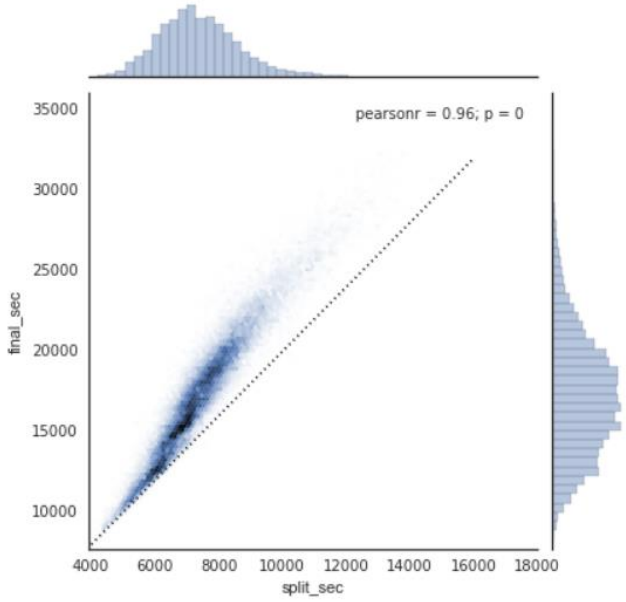
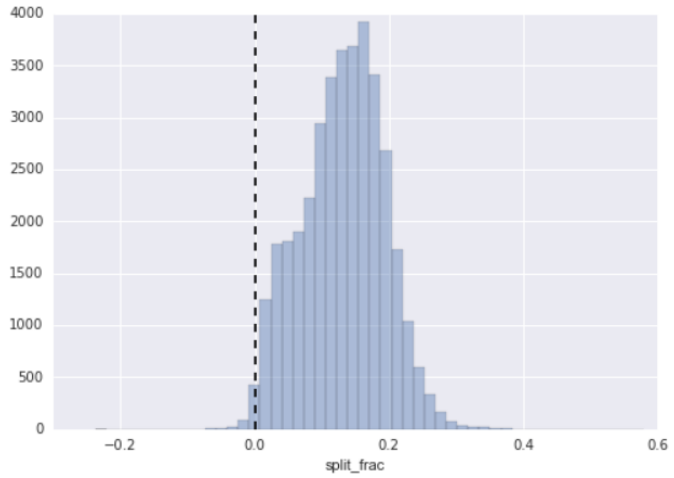
Data Science Assignment

This report is about the data visualization of a certain dataset to help understand the visualization process of the data science problem. The report discusses the dataset of a marathon race in which about 37000 male and female participants take part. The dataset comprises fields such as age, gender, start and end time of the race. Several aspects of the data are visualized. The data is 1st converted into numeric form and a joint plot is plotted as shown.

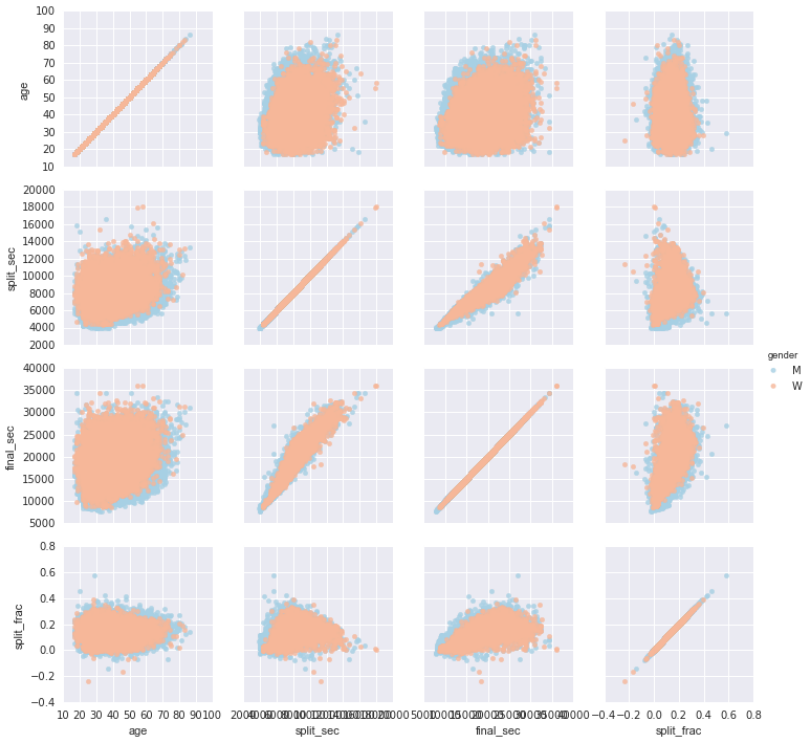


The dotted line in the graphs reflects the participant that has taken part in the marathon and kept a constant speed throughout. The dataset is further analyzed by introducing the split fraction which is either positive or negative value. Split fraction reflects either the runner goes positive or negative split. The negative value of the split fraction shows that the runner split the race negative by that value. The numeric result and bar graph showing the split fraction is passed below.

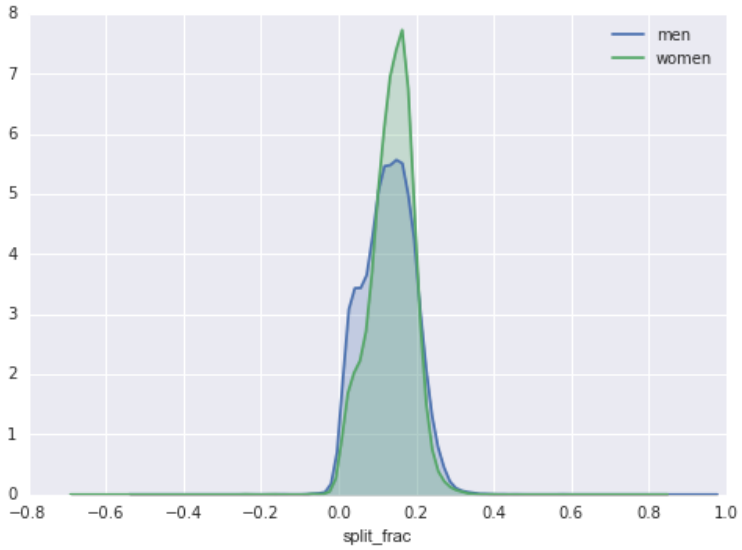
	age	gender	split	final	split_sec	final_sec	split_frac
0	33	M	01:05:38	02:08:51	3938.0	7731.0	-0.018756
1	32	M	01:06:26	02:09:28	3986.0	7768.0	-0.026262
2	31	M	01:06:49	02:10:42	4009.0	7842.0	-0.022443
3	38	M	01:06:16	02:13:45	3976.0	8025.0	0.009097
4	31	M	01:06:32	02:13:59	3992.0	8039.0	0.006842



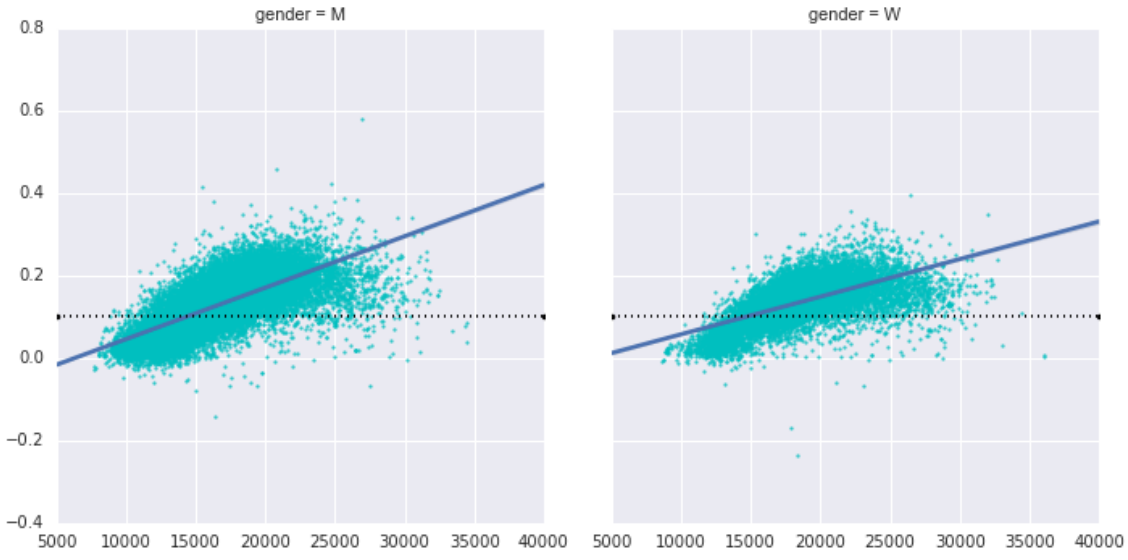
The bar graph shows that the total participants which were nearly 40000, only 250 participants were there who left the race in negative split. The negative split fraction is then combined with other features of the dataset to further analyze the dataset and its elements.



The pair plots above reflect that the split fraction has no relation with the age but has correlation with the time where a fast runner has closer to split on the marathon time. The correlation of the marathon with gender of the participant is also interesting as reflected by the histogram below.



The histogram reflects bimodal distribution between the genders. There are far more numbers of men than women in the race. Using the regression line for both male and female to reflect the relation of the gender with the split fraction, following results are obtained.



The regression line above reflects that runners with finishing time of 15000 seconds are the fast splits while those with slow rate of race are the second fast split.

Conclusion

The use of data visualization is very helpful in visualizing the data records, trending line and analysis as we found in the marathon case.