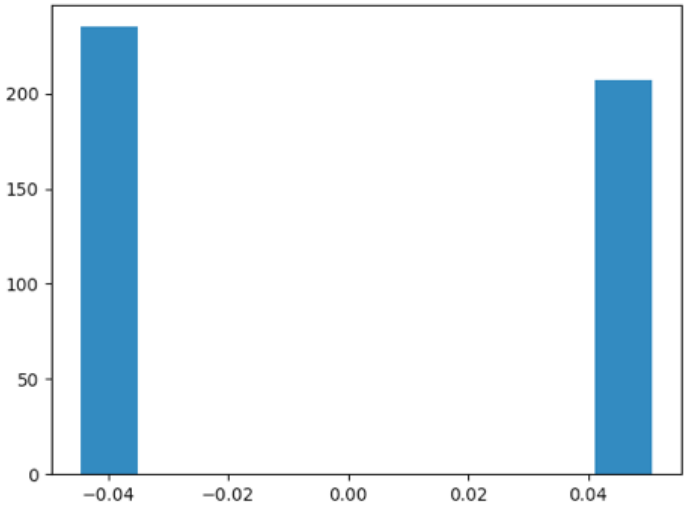
Assignment #8 – Diabates Dataset Analysis

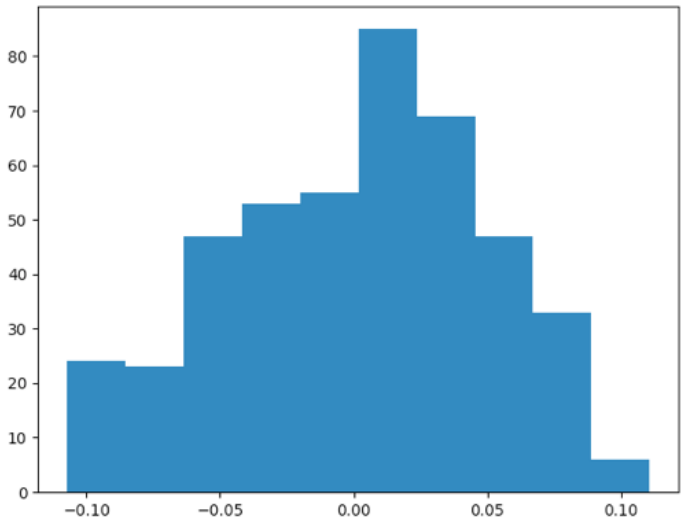
The diabetes dataset contains 442 instances, each with 10 attributes which include information such as age, sex, BMI, and multiple blood serum measurements. The target variable is "Interest," which measures the progression of diabetes after 1 year. In a clinical setting, such data would be useful in assessing risk and contributing factors to diabetes diagnoses, as well as potential preventative treatments.

1. Preliminary data exploration – histograms

By plotting histograms of the genders and ages of participants, we can see that there are a relatively even number of men and women in the study, as well as a roughly normal distribution in terms of age. In this way, we can say the dataset is generally balanced.



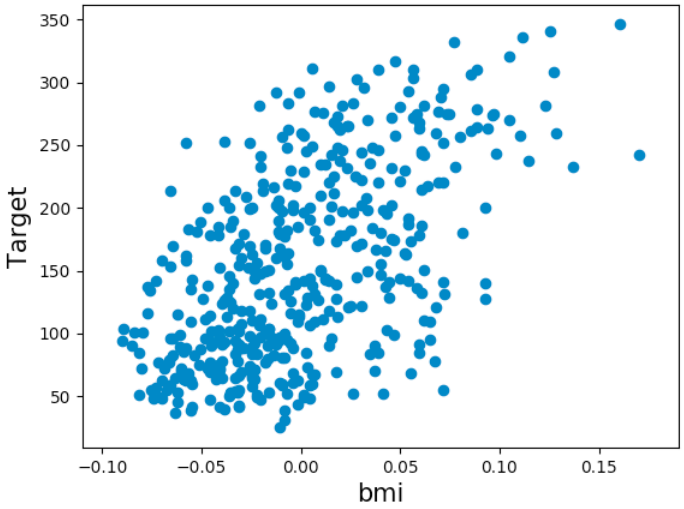
Occurrences by GENDER

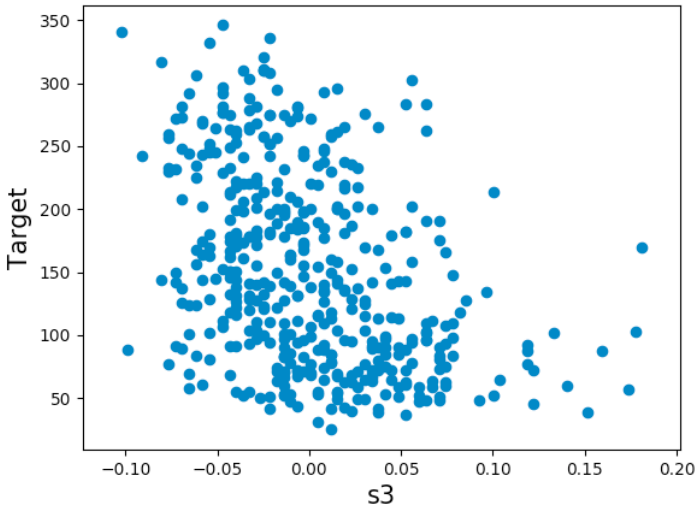


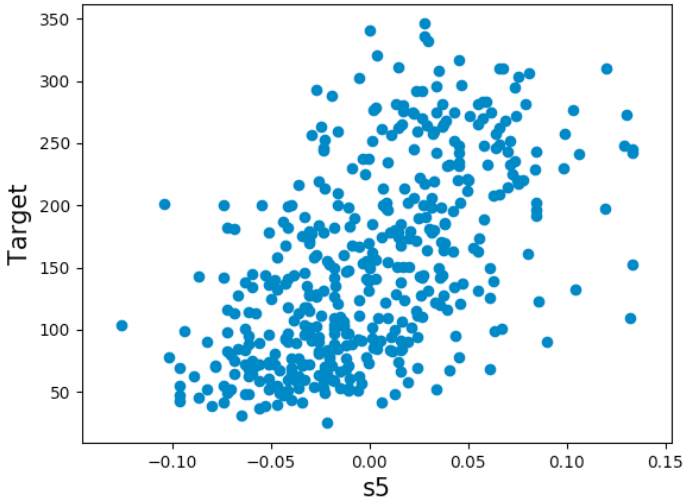
Occurrences by AGE

2. Preliminary scatter plots

With scatter plots of the attributes against the target variable we may conduct a visual assessment to see if there are visually notable or evident relationship. Initially, ‘BMI’, ‘s3’, and ‘s5’ all seem to be modest indicators of diabetes progression.





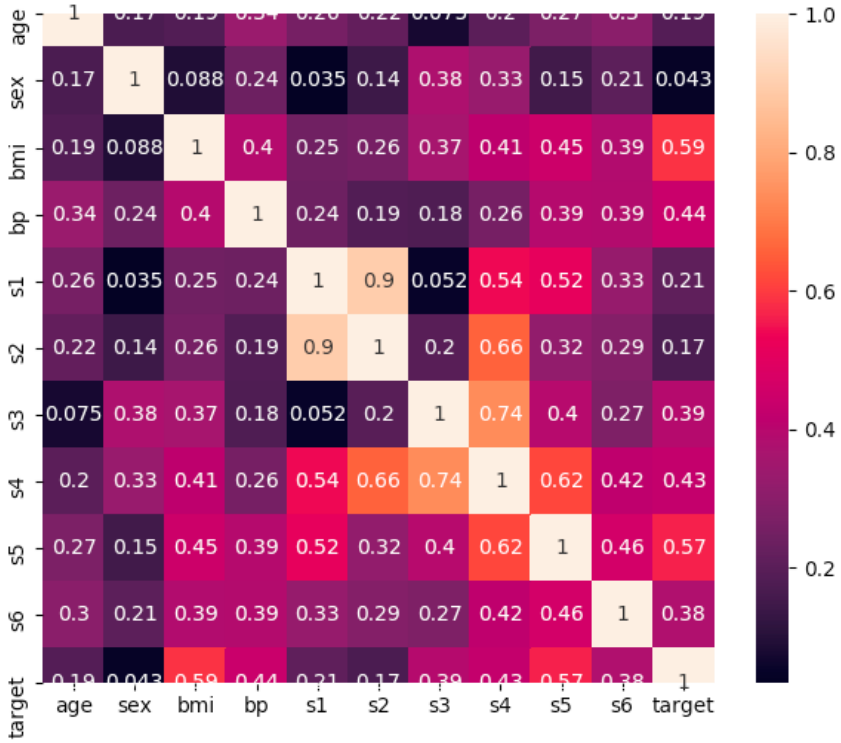


3. Produce a Heatmap to verify some of our previous assessments

Heatmaps are useful because they contain a lot of information at once. In this case, they test each attribute against the other attributes for correlation. In my code you can see that I also did this manually with the following code:

```
for feature_name in diabetes_data.feature_names:
    pearson = df.corr(method = 'pearson')
    corr_feature = pearson.loc[feature_name]
    print(corr_feature[abs(corr_feature).argsort()])
```

Interestingly, ‘s1’ and ‘s2’ are very predictive of one another (r = 0.9), as well as ‘s3’ and ‘s4’ (r = 0.74. For our target variable, ‘BMI’ and ‘s5’ carry the highest r-scores, which is in line with our visual assessment earlier. The relationship, however, is modest.



My recommendations for medical scientists would be to look into ‘BMI’ and ‘s5’ as indicators for diabetes progression and make recommendations to their clients based on these relations. Further, while I do not know what the different serums are, I would be interested in knowing if the high correlation between some of them are trivial or if they would be useful to examine.