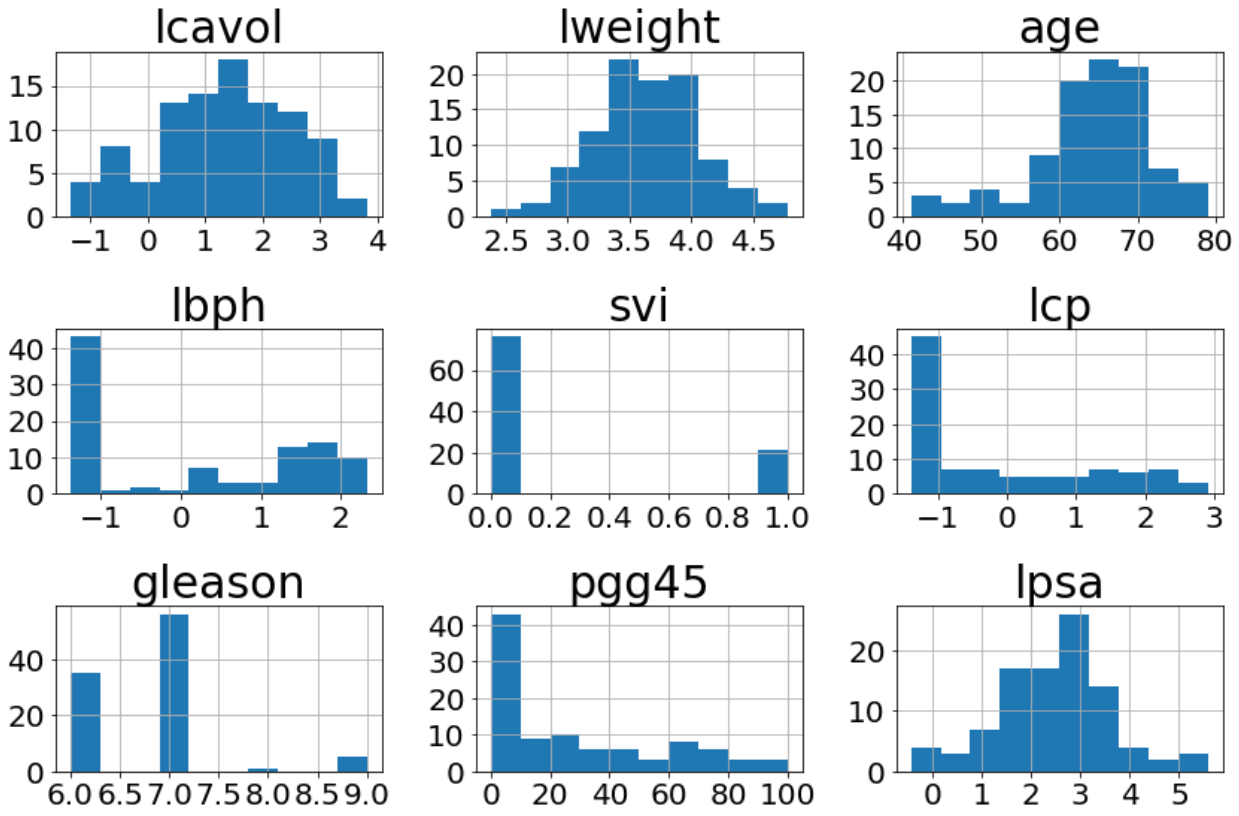
Rudy Dey
ECON 4850
Prostate Cancer Diagnosis Machine Learning Models
4/30/21

- Exploratory Data Analysis
  - Data Description

	lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45	lpsa
count	97.000000	97.000000	97.000000	97.000000	97.000000	97.000000	97.000000	97.000000	97.000000
mean	1.350010	3.628943	63.865979	0.100356	0.216495	-0.179366	6.752577	24.381443	2.478387
std	1.178625	0.428411	7.445117	1.450807	0.413995	1.398250	0.722134	28.204035	1.154329
min	-1.347074	2.374906	41.000000	-1.386294	0.000000	-1.386294	6.000000	0.000000	-0.430783
25%	0.512824	3.375880	60.000000	-1.386294	0.000000	-1.386294	6.000000	0.000000	1.731656
50%	1.446919	3.623007	65.000000	0.300105	0.000000	-0.798508	7.000000	15.000000	2.591516
75%	2.127041	3.876396	68.000000	1.558145	0.000000	1.178655	7.000000	40.000000	3.056357
max	3.821004	4.780383	79.000000	2.326302	1.000000	2.904165	9.000000	100.000000	5.582932

  - Predictor Variable Distribution



- Correlation Heatmap

Figure 1: Predictor Variable Correlation Matrix

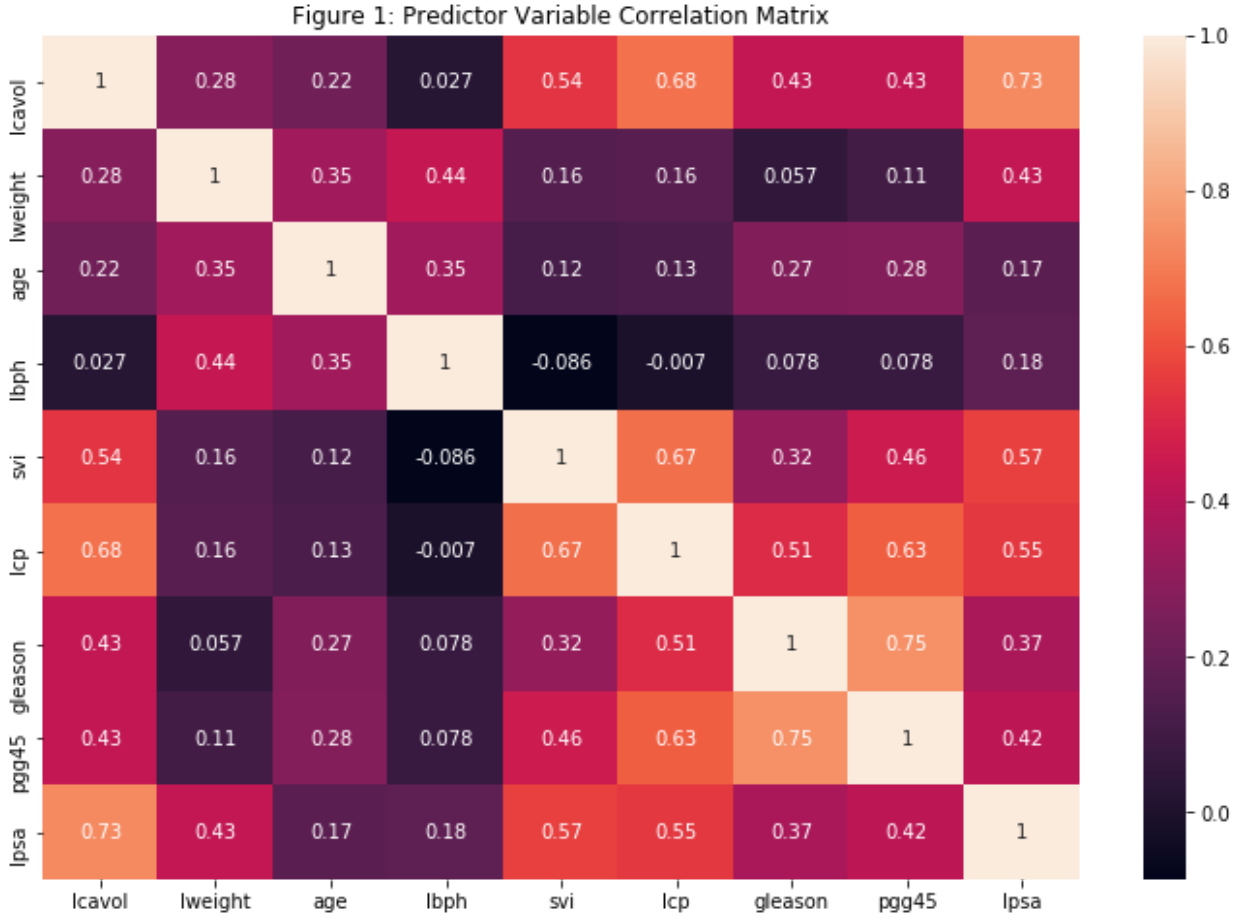
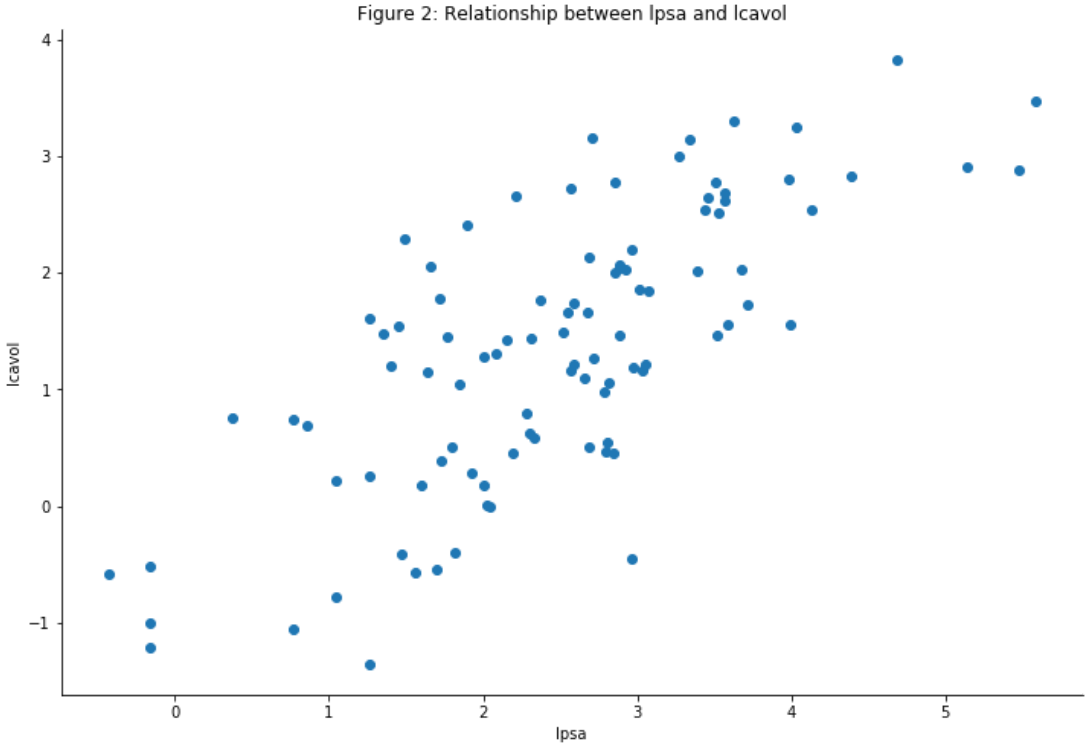


Figure 2: Relationship between lpsa and lcavol



The three key figures to take note of for the exploratory data analysis are the correlation heat map, the distribution graphs and the scatterplot between lcavol and lpsa.
The Predictor Variable Distribution graphs show that the outcome variable, along with lcavol, lweight and age, are all approximately normally distributed. This bodes well for the validity of the machine learning models to be developed.
The correlation heatmap of Figure 1 shows how strong correlated each of the variables are to each other and to the outcome variable, lpsa. From Figure 1, we see that the strongest correlations are between lpsa and lcavol, and pgg45 and gleason. The strong correlation between pgg45 and gleason suggests that there *may* be multicollinearity present, but it's not an issue at the moment. The correlation heatmap indicates which variables may have more importance in models like Random Forests: the stronger the correlation, the likelier that variable is in having a higher feature importance in a decision tree model.
The key takeaway from Figure 1 is that the predictor variable that has the strongest correlation with the outcome variable is lcavol. As such, Figure 2 is created to get a visualization of the relationship between lcavol and lpsa. Figure 2 shows that there is clearly a positive linear relationship between lcavol and lpsa.

- Machine Learning Using Linear Penalization Models
    - Each model was built after scaling the predictor variables using the standard scaler and splitting the dataset into training and testing datasets, with the testing dataset having 20% of the original dataset.
    - Table 1: Model Coefficient Estimates and Mean Squared Errors

	OLS	Ridge	Lasso	Elastic Net
lcavol	0.664409	0.664379	0.654965	0.654915
lweight	0.496889	0.496531	0.417405	0.417106
age	-0.026831	-0.026823	-0.023958	-0.023950
lbph	0.118506	0.118525	0.118480	0.118494
svi	0.783247	0.782134	0.609598	0.608736
lcp	-0.168766	-0.168471	-0.108863	-0.108622
gleason	0.138275	0.138063	0.057213	0.057078
pgg45	0.002174	0.002178	0.003424	0.003427
Training MSE	0.4365475217967732	0.4365476705696794	0.4419005952153945	0.4419424054804098
Testing MSE	0.5401691943434795	0.5400814088187056	0.5159025607326395	0.5158699585332689

Using the Testing MSE as the criterion, we can see that the Elastic Net model is the best model for finding the line of best fit. We can see that as the models become more complicated testing MSEs get smaller, which suggests the models become better at finding the line of best fit. The models can be improved by narrowing down the predictors variables and using only the best

ones, or by weighing the importance of each predictor variable in the model until the lowest MSE is found. Other, more complex models like Neural Network, SVG, or Gradient Vector may also show better MSE. Adding additional predictor variables can also improve MSE since we don’t know if there are other predictors that are not in the dataset that may be able better predict the lpsa.