CASE STUDY

PROSTATE CANCER

Content Outline

OVERVIEW OF PROSTATE CANCER

PROBLEM STATEMENT

DATA PRE-PROCESSING & CLEANING

EXPLORATORY DATA ANALYSIS

MODEL BUILDING

CONCLUSION

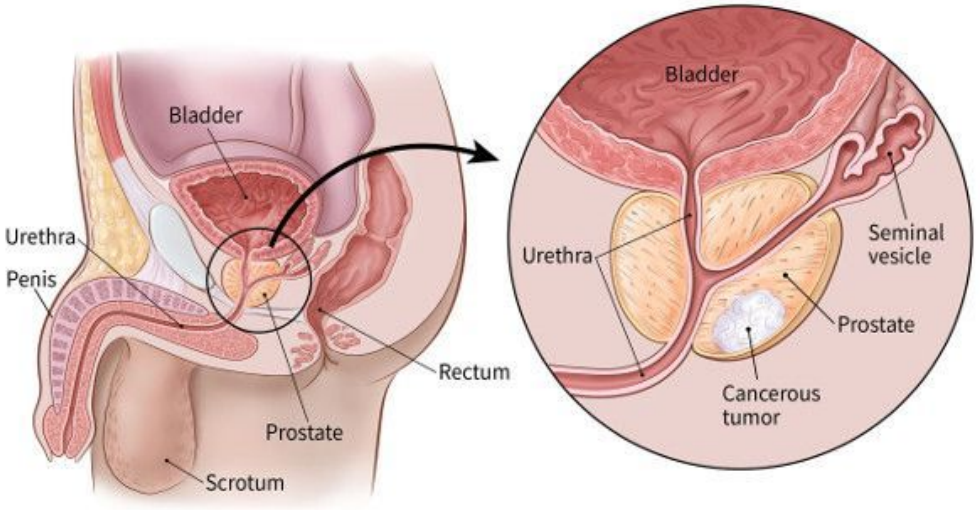
UNDERSTANDING PROSTATE CANCER

It develops in the prostate which is a small walnut shaped gland present in the pelvis of men.

Growths in the prostate can be benign (not cancer) or malignant (cancer).

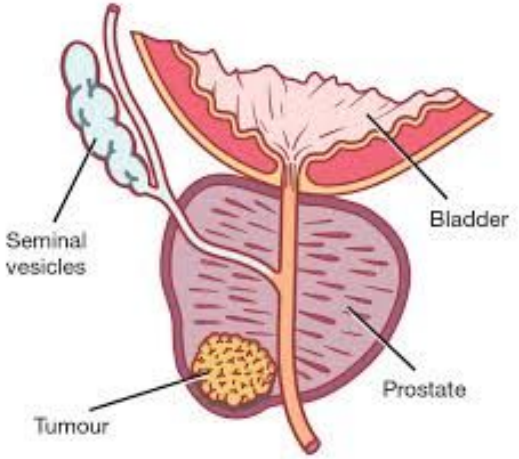
Benign prostatic hyperplasia (BPH) is when the prostate and surrounding tissue expands.

As men ages, they are more prone to prostate enlargement and are also more prone to the risk of prostate cancer.

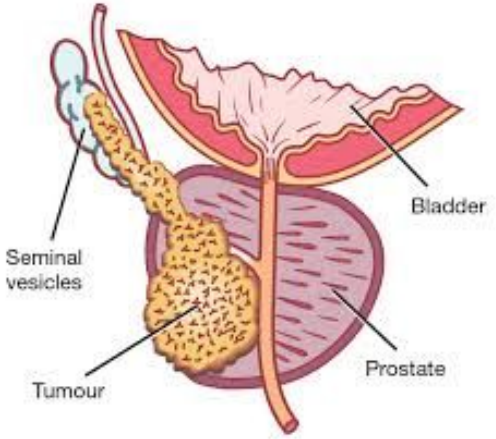


What is the PSA test ?

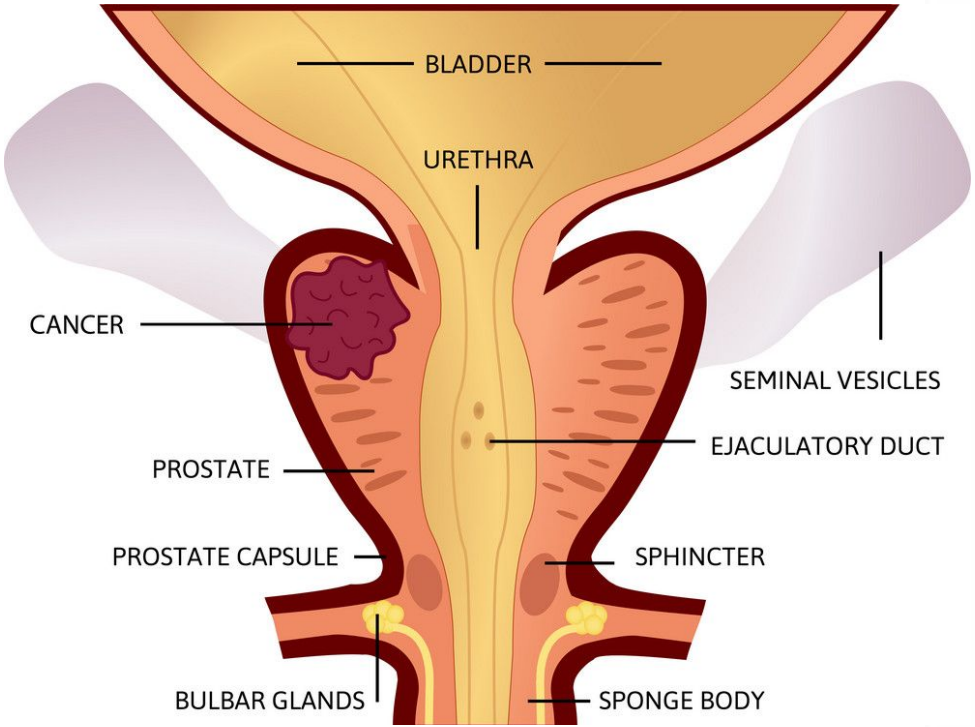
- Prostate-specific antigen (PSA) blood tests are used to screen for prostate cancer.
- A rapid rise in PSA may be a sign that something is wrong.
- If your prostate is enlarged or your PSA test comes back high, the abnormal results can be caused by prostate cancer

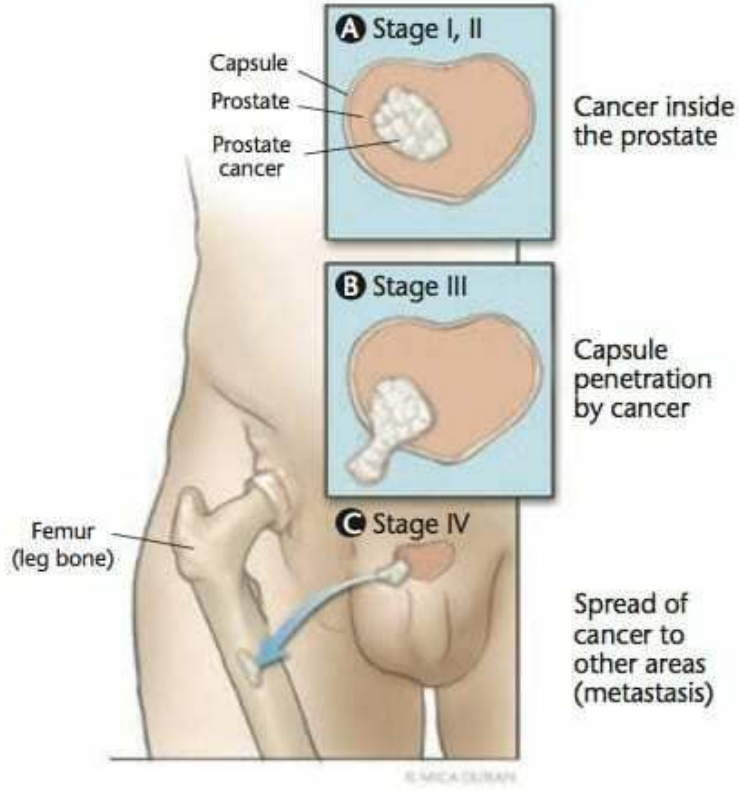


Seminal vesicle invasion



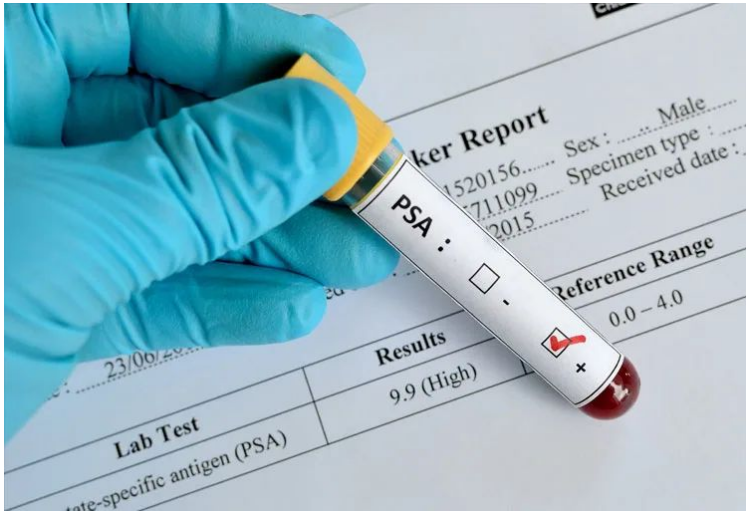
Capsular penetration





PROBLEM STATEMENT

To identify the predictors of psa which in turn aids in the determination of prostate cancer in patients.



VARIABLE DESCRIPTION

Total size : 97 x 9 ***Data file : prostate.txt***

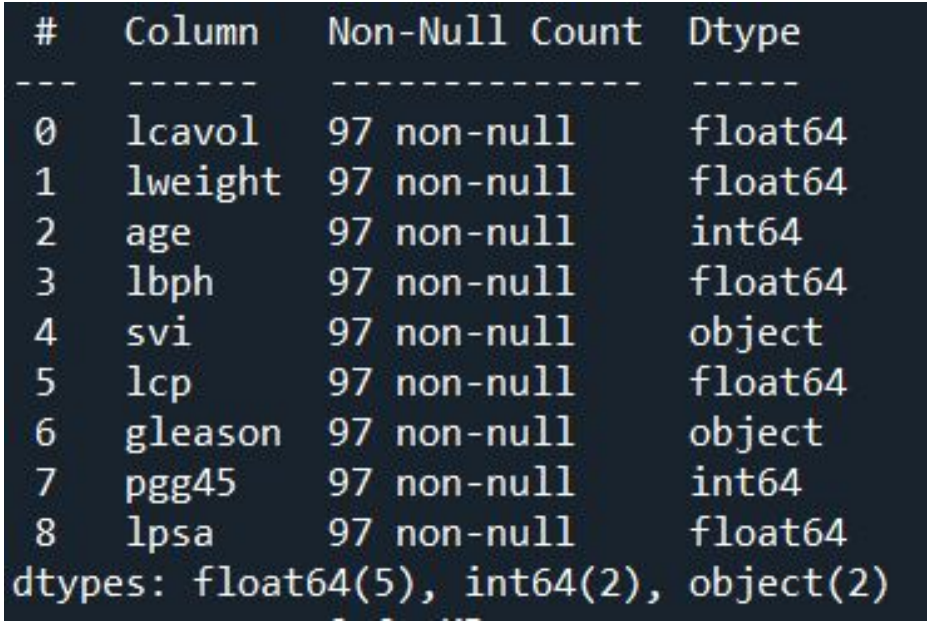
Variables	Description
lcavol	log cancer volume
lweight	log prostate weight
age	age
lbph	log of the amount of benign prostatic hyperplasia
svi	Seminal vesicle invasion
lcp	log of capsular penetration
gleason	Gleason score
pgg45	Percent of gleason scores 4 or 5
lpsa	log of prostate specific antigen

DATA SET

	lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45	lpsa	train
1	-0.579818	2.769459	50	-1.386294	0	-1.386294	6	0	-0.430783	
2	-0.994252	3.319626	58	-1.386294	0	-1.386294	6	0	-0.162519	
3	-0.510826	2.691243	74	-1.386294	0	-1.386294	7	20	-0.162519	
4	-1.203973	3.282789	58	-1.386294	0	-1.386294	6	0	-0.162519	
5	0.7514161	3.432373	62	-1.386294	0	-1.386294	6	0	0.3715636	
6	-1.049822	3.228826	50	-1.386294	0	-1.386294	6	0	0.7654678	
7	0.7371641	3.473518	64	0.6151856	0	-1.386294	6	0	0.7654678	
8	0.6931472	3.539509	58	1.5368672	0	-1.386294	6	0	0.8544153	
9	-0.776529	3.539509	47	-1.386294	0	-1.386294	6	0	1.047319	
10	0.2231436	3.244544	63	-1.386294	0	-1.386294	6	0	1.047319	
11	0.2546422	3.604138	65	-1.386294	0	-1.386294	6	0	1.2669476	
12	-1.347074	3.598681	63	1.2669476	0	-1.386294	6	0	1.2669476	
13	1.6134299	3.022861	63	-1.386294	0	-0.597837	7	30	1.2669476	
14	1.4770487	2.998229	67	-1.386294	0	-1.386294	7	5	1.3480731	
15	1.2059708	3.442019	57	-1.386294	0	-0.430783	7	5	1.3987169	
16	1.5411591	3.061052	66	-1.386294	0	-1.386294	6	0	1.446919	
17	-0.415515	3.516013	70	1.2441546	0	-0.597837	7	30	1.4701758	

DATA PREPROCESSING

- Column named ‘train’ with no values was dropped.
- No missing values were found.
- No duplicate records.
- Data type of column ‘svi’ and ‘gleason’ was changed to object as it contained categorical values.



#	Column	Non-Null Count	Dtype
0	lcavol	97 non-null	float64
1	lweight	97 non-null	float64
2	age	97 non-null	int64
3	lbph	97 non-null	float64
4	svi	97 non-null	object
5	lcp	97 non-null	float64
6	gleason	97 non-null	object
7	pgg45	97 non-null	int64
8	lpsa	97 non-null	float64

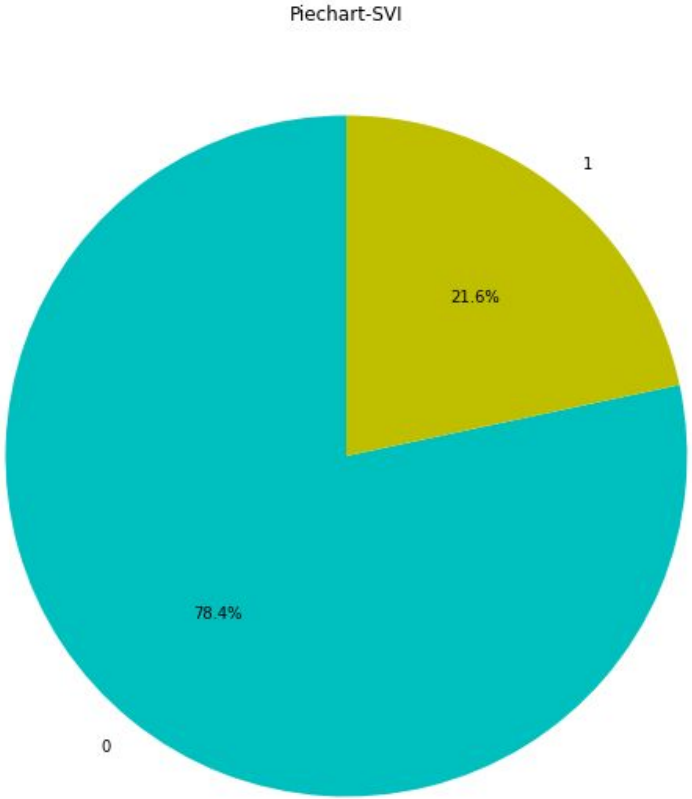
dtypes: float64(5), int64(2), object(2)

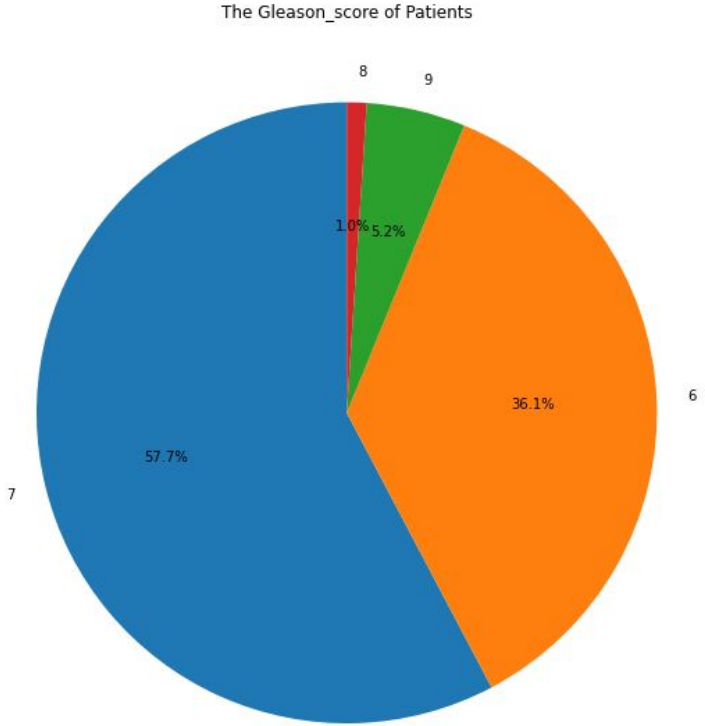
Summary of the data:

Index	lcavol	lweight	age	lbph	lcp	pgg45	lpsa
count	97	97	97	97	97	97	97
mean	1.35001	3.62894	63.866	0.100356	-0.179366	24.3814	2.47839
std	1.17862	0.428411	7.44512	1.45081	1.39825	28.204	1.15433
min	-1.34707	2.37491	41	-1.38629	-1.38629	0	-0.430783
25%	0.512824	3.37588	60	-1.38629	-1.38629	0	1.73166
50%	1.44692	3.62301	65	0.300105	-0.798508	15	2.59152
75%	2.12704	3.8764	68	1.55814	1.17866	40	3.05636
max	3.821	4.78038	79	2.3263	2.90417	100	5.58293

- The target variable ‘lpsa’ and ‘lcavol’ is slightly skewed to the left.
- ‘lweight’ shows normal distribution.
- Age is also skewed to the left with more than 50% of the males being 65 or above.
- Lbph shows left skewness as median> mean.

DISTRIBUTION OF CATEGORICAL VARIABLES





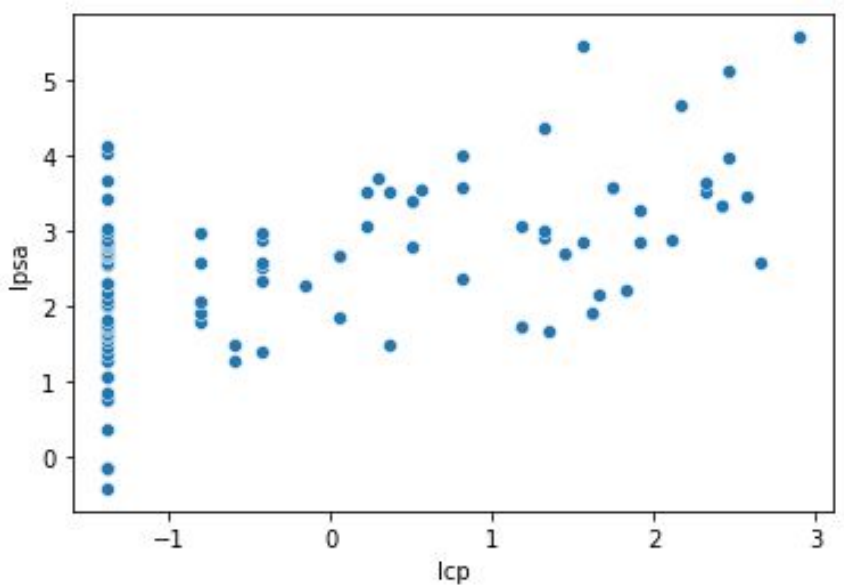
DATA VISUALISATION

Numerical Variables against target variable

Lcp vs lpsa

- We can observe an increase in lpsa value with increase in lcp.

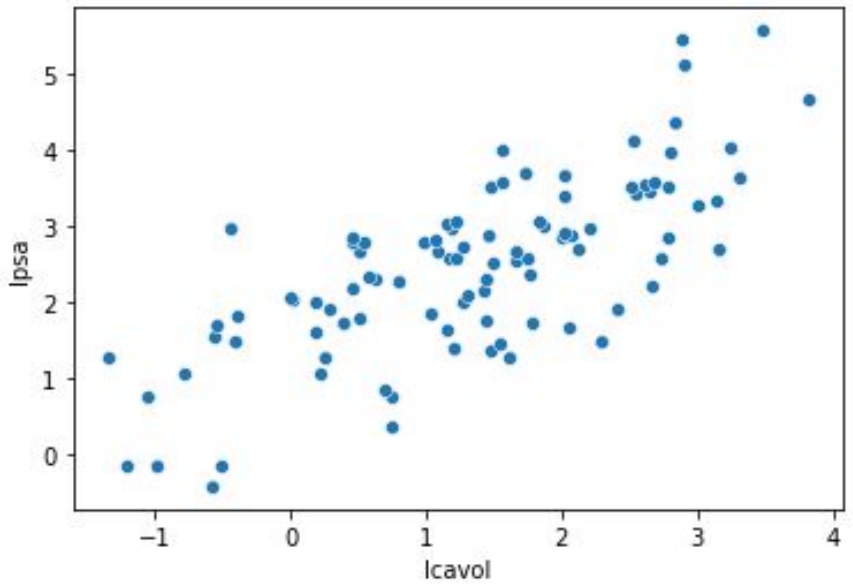
- This suggest that the lcp is highly influences the lpsa score denoting the presence of cancer.



Lcavol vs lpsa

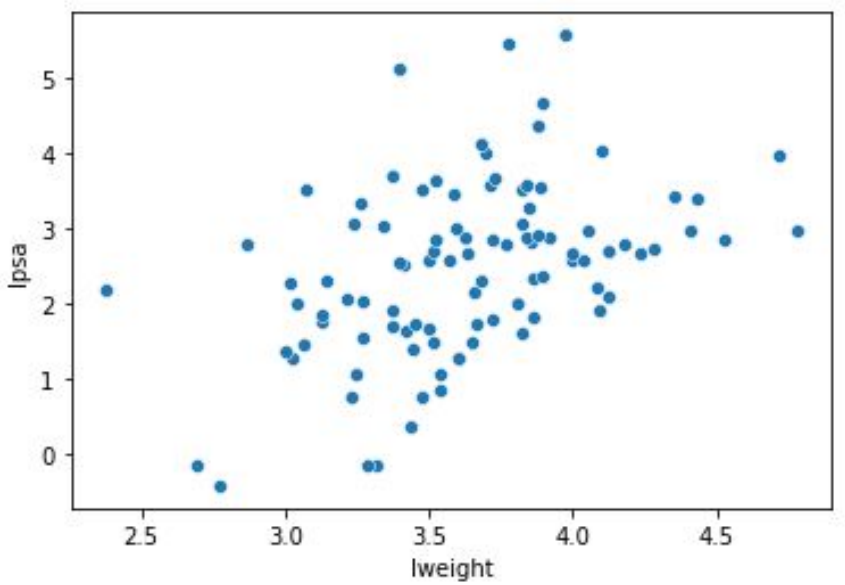
- We can observe a clear trend where increase in lcavol showed an increase in the lpsa values.

- This suggests that the cancer volume seemed to have a significant impact on lpsa.



Lweight vs lpsa

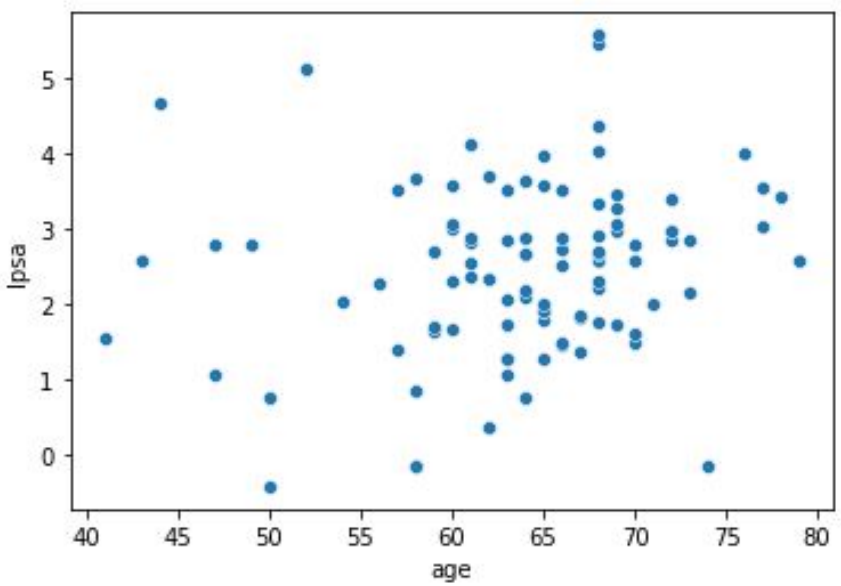
- There is no clear trend for us to conclude that the weight of the prostate gland alone determines the value of the psa test.
- However we can see a slight increase in psa values as weight of the prostate gland increases.



Age vs lpsa

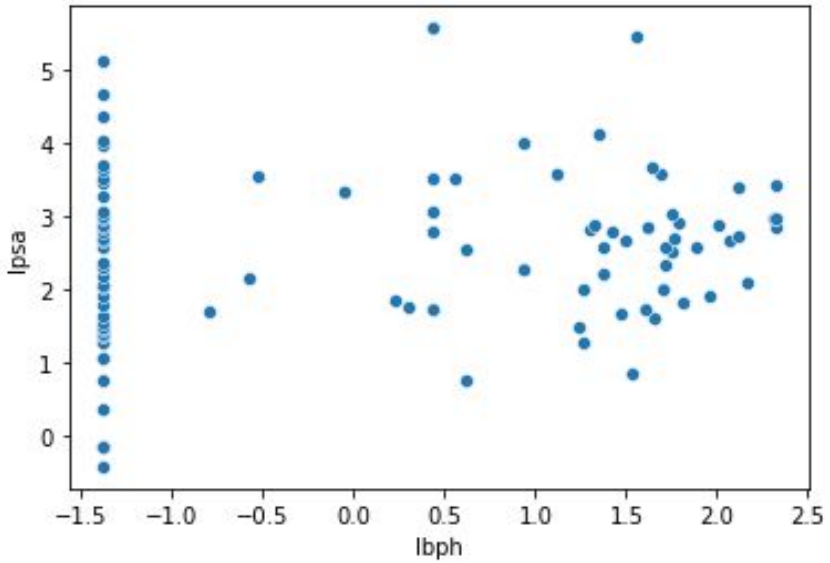
- Most of the men in the dataset are in the range 60-70 years.

- In older men, we can see a slight increase in the value of psa when compared to younger men where there appears to be no trend.



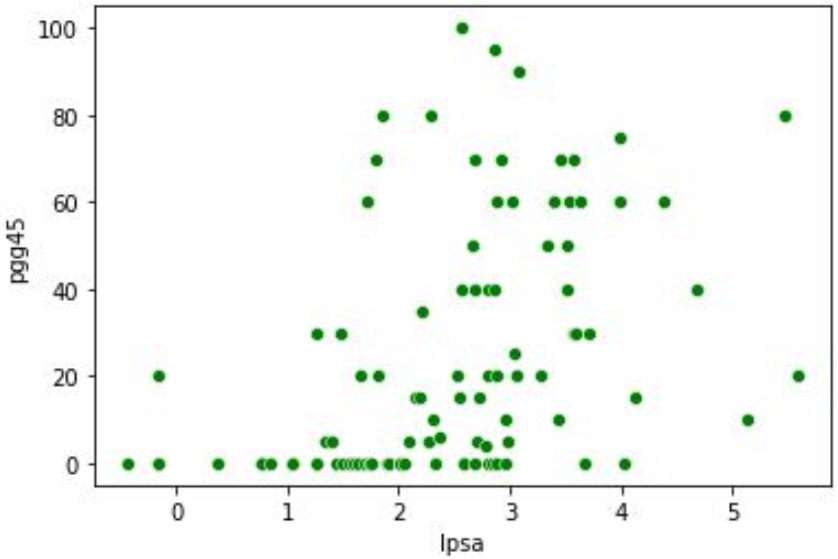
Lbph vs lpsa

There is no clear pattern or variation observed so it can be inferred that the variable lbph does not affect lpsa.



Pgg45 vs lpsa

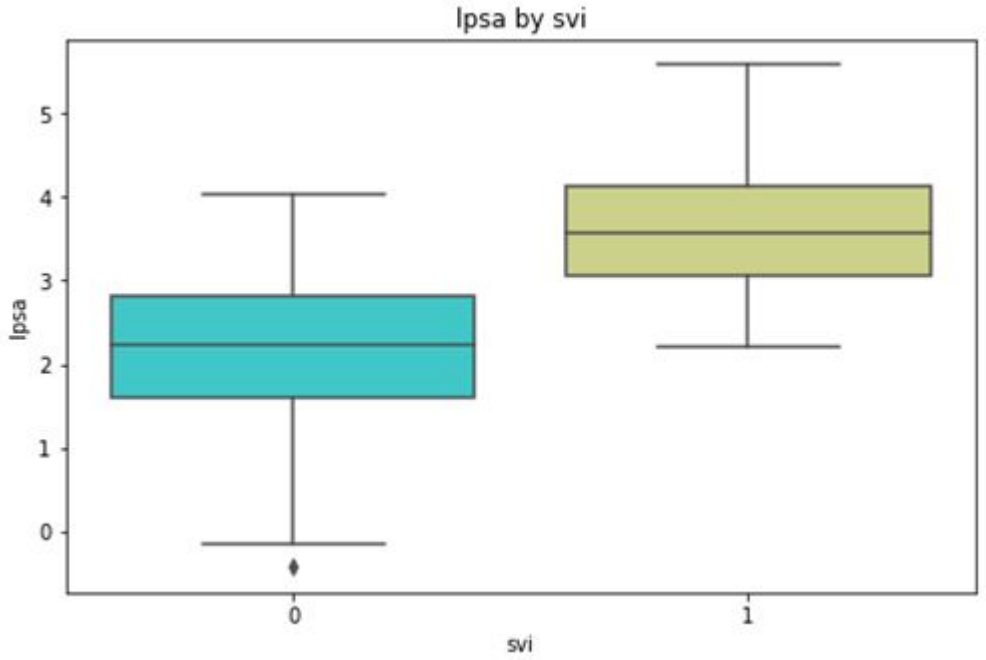
The variable pgg45 does not show a clear pattern or variation with lpsa so there seems to be a weak correlation between pgg45 and lpsa.



Categorical variables against lpsa

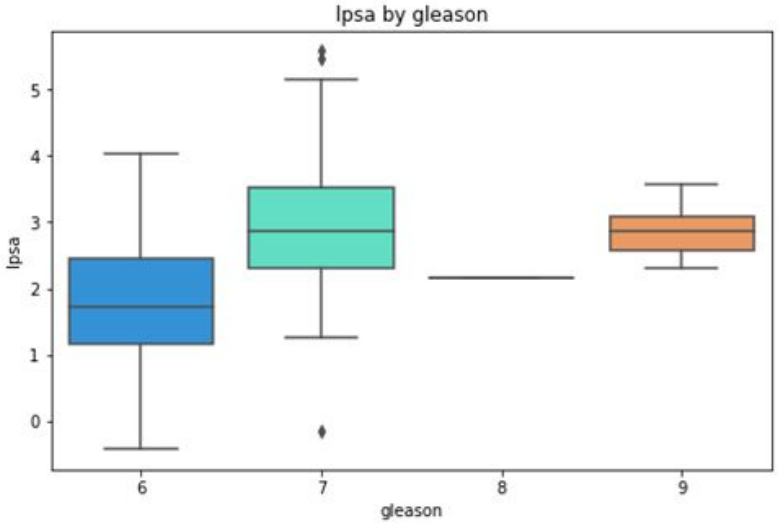
SVI vs lpsa

- Median of lpsa value is higher for men having svi
- 78.3% of the men do not have svi.The remaining 21.64% of the men with svi, show an median lpsa value of 3.5 .
- People with svi have higher prostate specific antigen



Gleason vs lpsa

- The Gleason score tells us how much the cancer cells look like normal cells.
- We can see an inconsistent increase in lpsa value as gleason value increases with gleason score 8 having negligible count .
- We cannot say it alone determines lpsa.

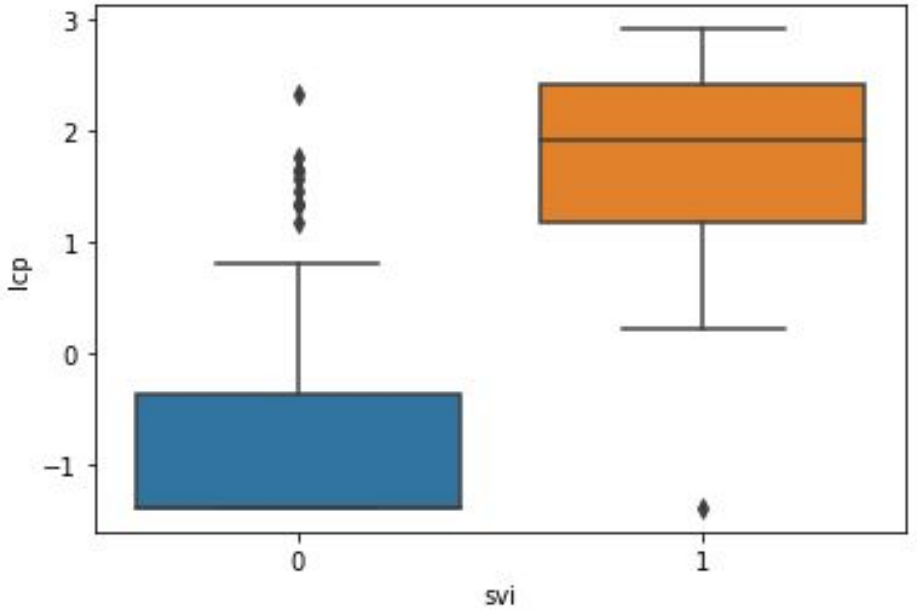


Index	gleason
6	35
7	56
8	1
9	5

Plots between independent variables

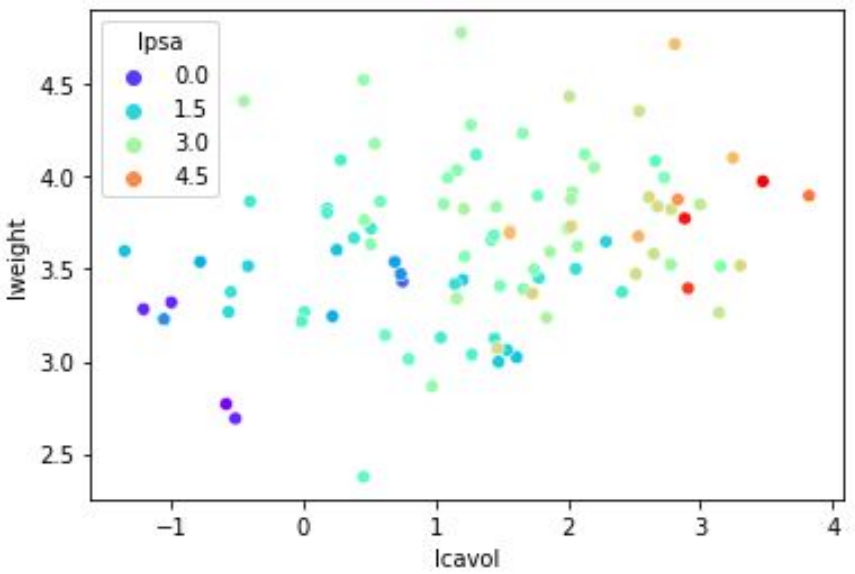
SVI vs lcp

- Patients with svi have a higher value of lcp which coincides with the fact that presence of cancer is seen with people having high lcp.



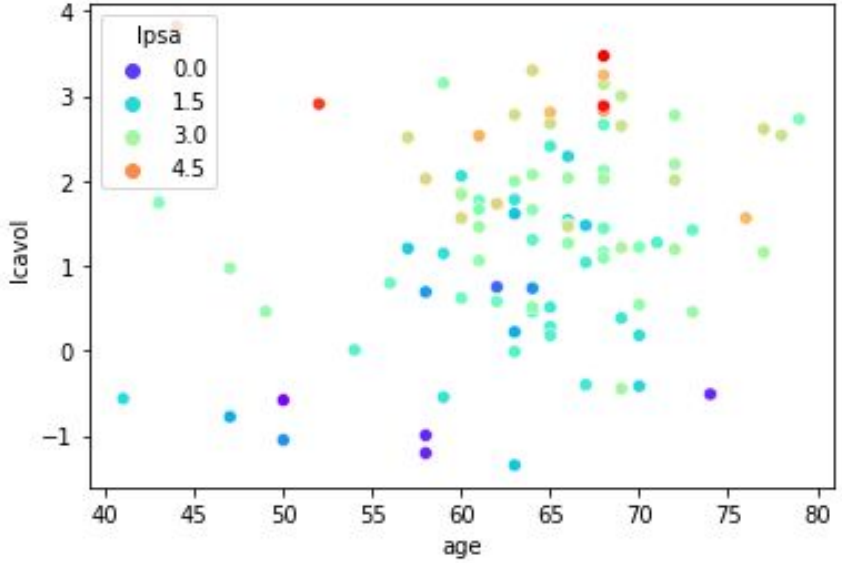
Scatter plot of lcavol and lweight and lpsa

- We can’t see a clear trend where it suggests that the cancer volume depends on the weight of the patient’s prostate.
- However, we can see that lpsa of the patient is increasing with the volume of cancer .This suggests that lpsa shows strong dependency on volume of the cancer.



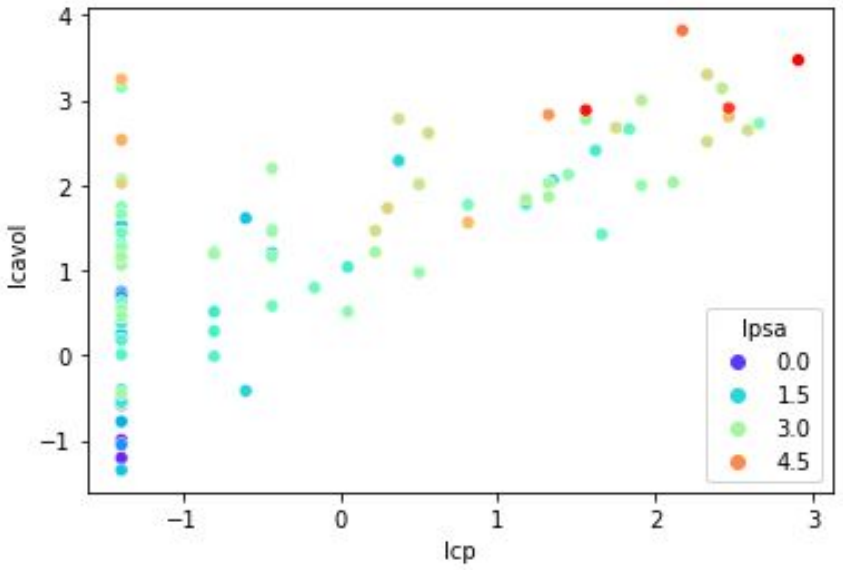
Scatter plot between age and lcavol with lpsa

- We can see that majority of the men with a higher cancer volume belong in the age group of 60-70.
- It is also observed that lpsa is considerably high in men with a higher cancer volume.
- We can conclude that age of the patient do affect the lpsa value to some extent as younger men have lpsa value in the lowest range.

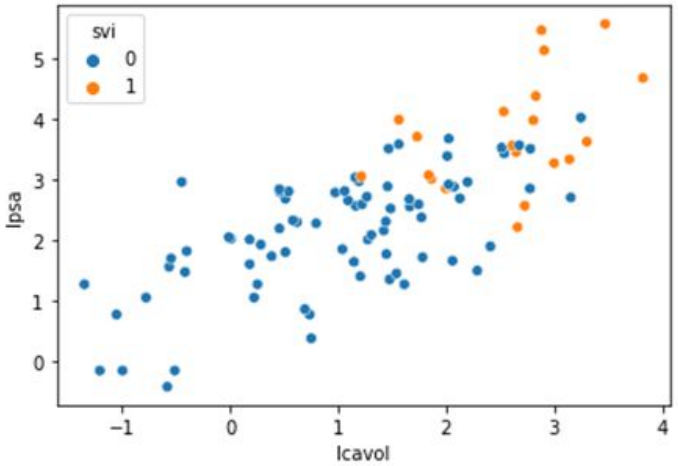


Scatter plot between lcavol and lcp with lpsa

- We can see a clear trend where there is an increase in volume of the cancer as log of capsular penetration increases.
- In addition to that, we can see an increase in the lpsa value with increase of those two variables.
- This shows a strong relationship between those two variables and with that of the target variable.
- This suggests the possibility of multicollinearity between these variables.

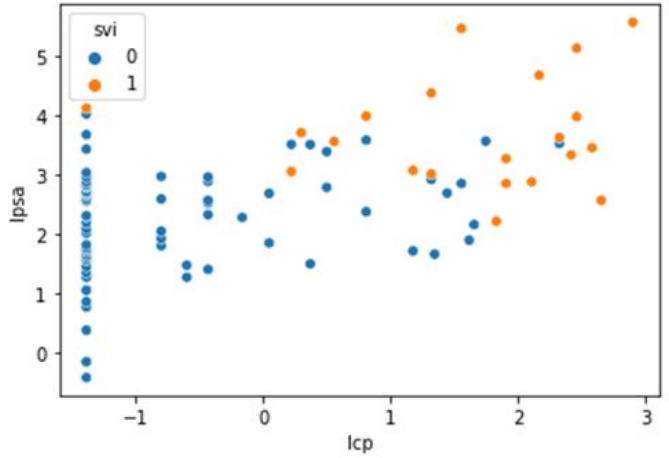


Lpsa vs lcavol and SVI



- People having high lcavol and lpsa tend to have svi value of 1.There's a clear trend in this comparison
- All the people who have lower cancer volume and lpsa don't have invasion of seminal vesicles.

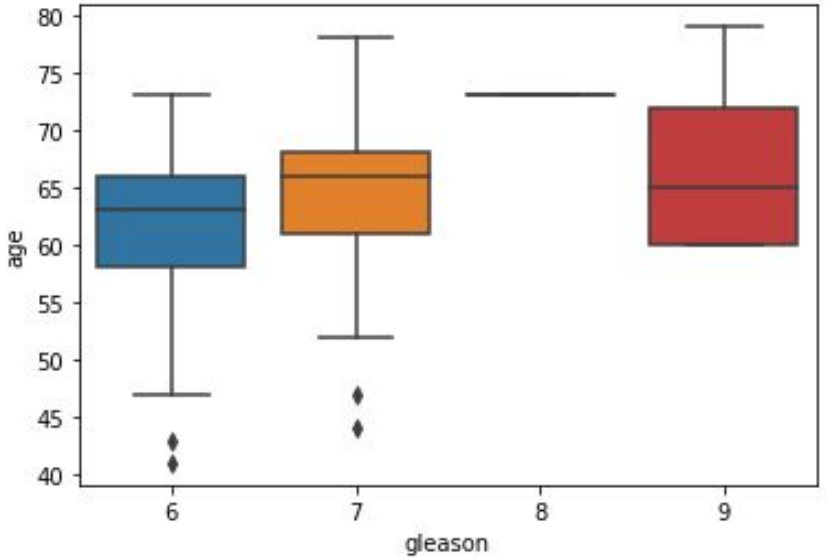
Lpsa vs lcp and SVI



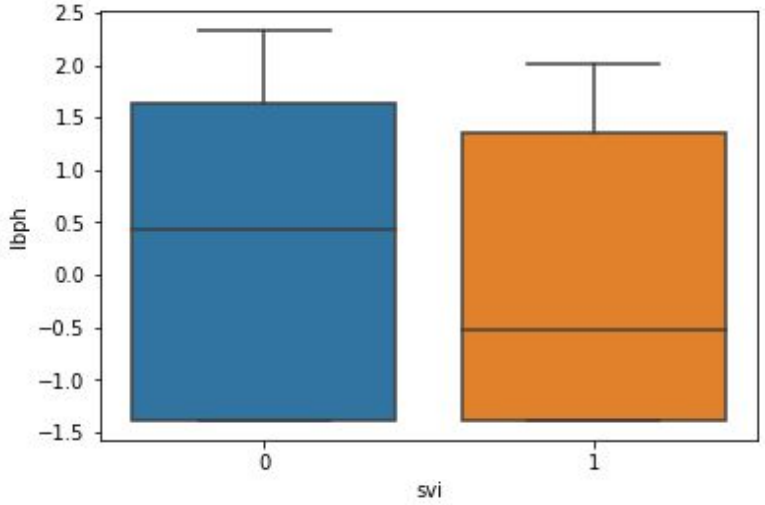
- People having higher capsular penetration and lpsa tend to have svi.

Gleason vs age

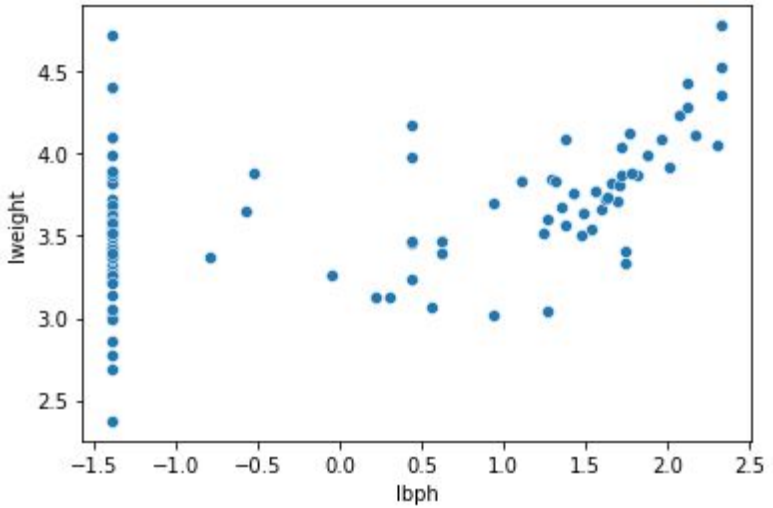
We can also see an increase in gleason score with age which denotes that prostate enlargement is seen as men age which can denote the presence of cancer or benign tissue.



lbph vs svi



lbph vs lweight



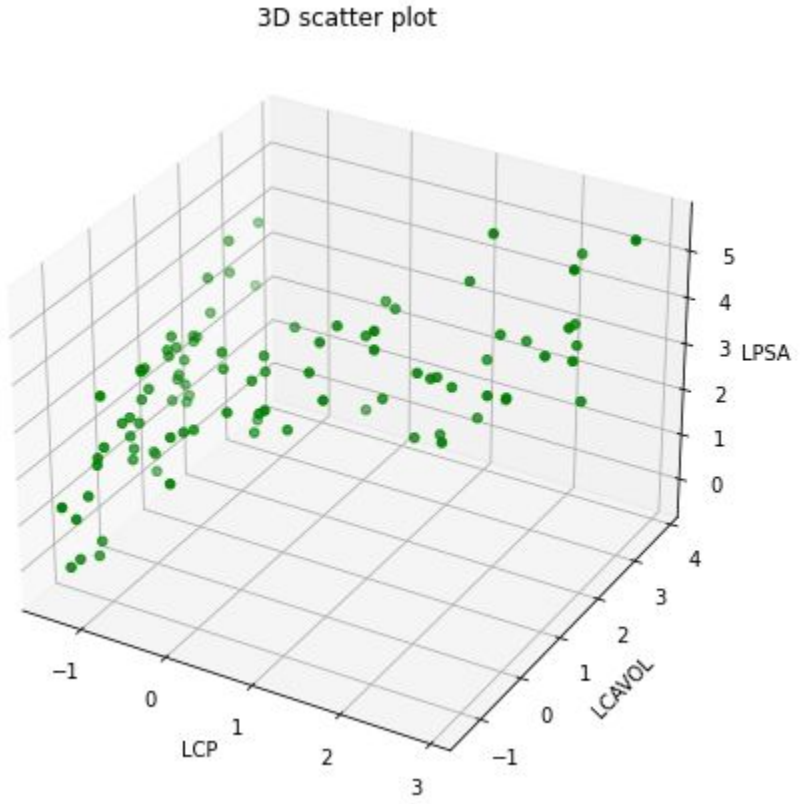
- Benign prostate hyperplasia which is prostate enlargement is a common condition in older men. This is seen in the scatterplot were lbph increases with prostate weight.

- Through the boxplot we can see that 78% of men which has **no svi has a higher lbph value** which confirms our earlier plots that suggested men with cancer have svi otherwise **they have enlarged prostate due to benign tissue.**

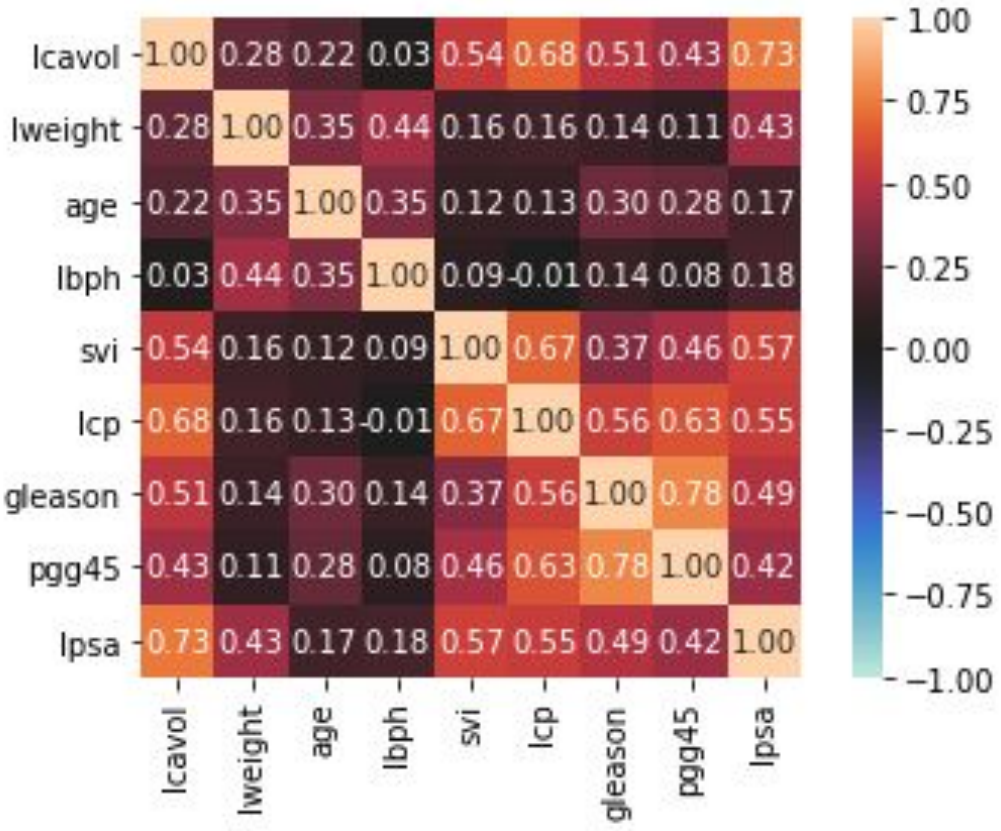
3D SCATTER PLOT

- Both lcp and lcavol increases with the values of lpsa

- Hence lcp and lcavol shows strong correlation with lpsa



CORRELATION



- ➔ The target variable lpsa shows strong correlation with variables lcavol,svi and lcp
- ➔ Age and lbph shows no correlation with lpsa
- ➔ Multicollinearity is observed between variables

SUMMARY :

- We saw a strong positive dependency between lcp and lcavol with each other and with our target variable.
- It was also observed that men with svi showed higher values for lpsa.
- As men got older, their gleason score showed an increase due to enlargement of their prostate which can either signify bph or malignant cancer.
- Lbph scores were higher with men with no svi which tells us that svi positively impacts the psa score denoting the grade of cancer.

MODEL BUILDING

MODEL BUILDING

MULTIPLE LINEAR REGRESSION (MLR)

A multiple linear regression model was built including all the variables in the data set.

We obtain a RMSE value of 0.7914 with a R^2 value of 53.83%

```
Root Mean Square Error: 0.7914738807781055
```

```
R2 value: 53.835134789978O8 %
```

MLR after dropping insignificant variables

- Dropped variables: lbph,age,pgg45

- We obtain a RMSE value of 0.7865 with a R^2 value of 55.950%

```
Root Mean Square Error: 0.7865001996669052
```

```
R2 value: 55.95014166053629 %
```

MODEL SELECTION

The multiple linear regression model built after dropping variables such as pgg45, lbph & age showed higher r2 value and a lower RMSE value compared to model including all the features so choosing this model to predict lpsa will be appropriate.

CONCLUSION

From the above analysis, it can be concluded that variables such as lcavol(log of cancer volume), svi(seminal vesicle invasion), lcp (log of capsular penetration), lweight(log prostate weight), gleason score act as predictors of lpsa (log of prostate specific antigen) which in turn can be used to predict prostate cancer in patients.

THANK YOU