Predicting Prostate-Specific Antigen (PSA) Levels in Men with Advanced Prostate Cancer

Olin Yoder

List of Tables

List of Figures

Title

Predicting Prostate-Specific Antigen (PSA) Level in Men with Advanced Prostate Cancer

Abstract

This analysis explores the relationship between various prognostic clinical measurements and their relationship with Prostate-Specific Antigen (PSA) levels in men with advanced prostate cancer. The variables in the dataset are cancer volume, weight, age, hyperplasia, seminal vesicle invasion, capsular penetration, and Gleason score. The primary goal is to identify which of these variables are key predictors of PSA and assess how well these variables explain its variation. Multiple regression techniques, including linear, ridge, and lasso regression, are applied to develop predictive models and evaluate their performance.

Introduction

This analysis focuses on explaining the relationships between various prognostic clinical measurements in men with advanced prostate cancer and Prostate-Specific Antigen (PSA) levels, with the goal of developing a predictive model for PSA. The dataset includes 97 observations across nine variables: id, cancer volume (*cancerv*), weight, age, hyperplasia, seminal vesicle invasion (*seminal*), capsular penetration (*capsular*), and Gleason score (*score*). Various multiple regression techniques were employed, including basic multiple linear regression, ridge regression, and lasso regression. Categorical variables, *seminal* and *score*, were dummy encoded for inclusion in the models. Model performance was evaluated using adjusted R^2, RMSE, and 10-fold cross-validation. The results show the majority of variables contribute to explaining PSA levels, with seminal vesicle invasion, Gleason score, and cancer volume being consistently significant predictors across models. In terms of model performance, ridge regression performed similarly to other models in terms of explanatory power but offered improved generalization on unseen data. Despite the moderately-strong predictive power of the models, the small sample size of the dataset (97 observations) is limiting in generalizing the model to a greater population. Future analysis with larger datasets may provide more reliable results.

Primary Analysis Objective

To investigate the relationship between various prognostic clinical measurements in men with advanced prostate cancer to predict their PSA level.

Materials & Methods

Data Sources

The data set was adapted in part from: *Hastie, T. J.; R. J. Tibshirani; and J. Friedman. The Elements of Statistical Learning: Data Mining. Inference. and Prediction. New York: Springer-Verlag, 2001*. The data set features 97 observations across nine variables, including an ID field. The names of the nine variables are: *idnum*, *psa*, *cancerv*, *weight*, *age*, *hyperplasia*, *seminal*, *capsular*, and *score*. Further information about the variables can be found in the data dictionary on page 6. The data set contains no null values.

Statistical Analysis

The data is provided in .xlsx format. All analyses were conducted using R (version 4.2.1). The primary objective of the analysis is to identify relationships between the given variables and PSA levels, specifically to determine how these variables can predict PSA levels. To achieve this, multiple linear regression was selected as the modeling approach. Before training any regression model, each variable was examined individually to understand its relationship with PSA levels. Next, various multiple regression techniques were then applied, ultimately leading to a final model and the interpretation of said model.

Model Assumptions

Unless otherwise noted, all inferences are made using a significance level (α) of 0.05. Two variables, *seminal* and *score*, are categorical and are dummy encoded during the regression process. Additionally, model assumptions, including the constancy of error variances, normality of the error terms, independence of errors, and linearity of the relationship between the outcome and predictors, were verified before finalizing the estimated fitted regression model.

Primary Objective Analysis

To explore each variable and its relationship with PSA levels. Then, use the variables that best explain PSA levels to predict PSA levels and assess how well these variables were able to predict PSA levels.

Summary of Dataset

Table 1, below, contains the data dictionary.

Variable	Description
IDNum	Identification number
PSA	Prostate-specific antigen level (mg/ml)
Cancerv	Cancer volume (cc)
Weight	Prostate weight (gm)
Age	Age (years)
Hyperplasia	Amount of benign prostatic hyperplasia (cm^2)
Seminal	Presence or absence of seminal vesicle invasion (1 = yes, 0 = otherwise)
Capsular	Degree of capsular penetration (cm)
Score	Gleason score, pathologically determined grade of disease using total score of two patterns (summed scores were either 6, 7, or 8, with higher scores indicating worse prognosis)

Table 1: Data Dictionary

Table 2, below, contains five figure summary statistics of the PSA dataset.

Descriptor	IDNum	PSA	CancerV	Weight	Age	Hyperplasia	Capsular	Seminal	Score
Min.	1	0.651	0.2592	10.70	41.00	0.000	0.0000	0.0000	6.000
1st Qu.	25	5.641	1.6653	29.37	60.00	0.000	0.0000	0.0000	6.000
Median	49	13.330	4.2631	37.34	65.00	1.350	0.4493	0.0000	7.000
Mean	49	23.730	6.9987	45.49	63.87	2.535	2.2454	0.2165	6.876
3rd Qu.	73	21.328	8.4149	48.42	68.00	4.759	3.2544	0.0000	7.000
Max.	97	265.072	45.6042	450.34	79.00	10.278	18.1741	1.0000	8.000

Table 2: Summary Statistics

Variable Analysis

PSA (Prostate-Specific Antigen) Levels

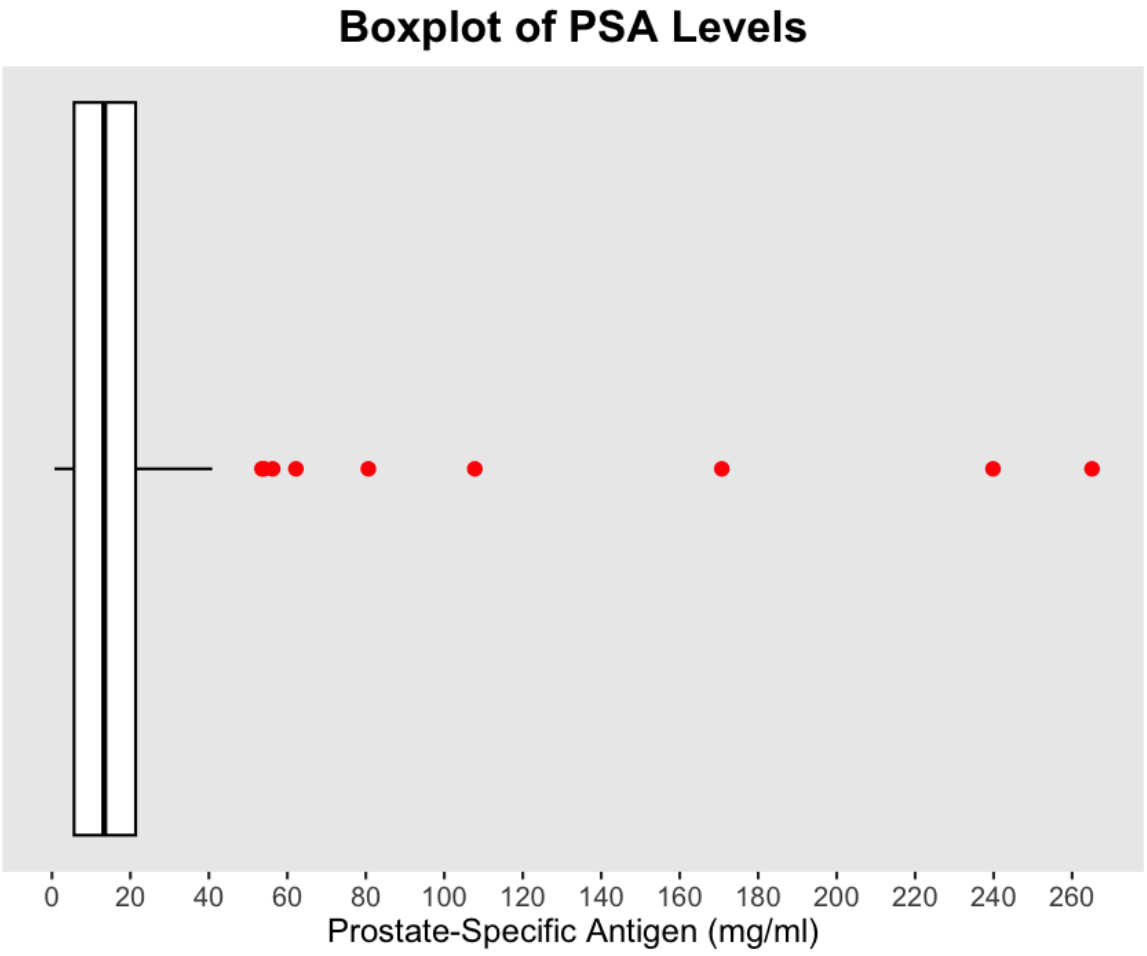

Figure 1: PSA Level

Figure 1 shows the distribution of PSA readings. The graph shows the median values is concentrated around 13 mg/ml, with the interquartile range (IQR) spanning from approximately 5 mg/ml to just above 20 mg/ml. However, the distribution appears to be heavily right-skewed, with a few extreme values. Notably, there are three outliers with PSA values exceeding 150 mg/ml, nearly twice as high as the fourth highest reading.

PSA vs Cancer Volume

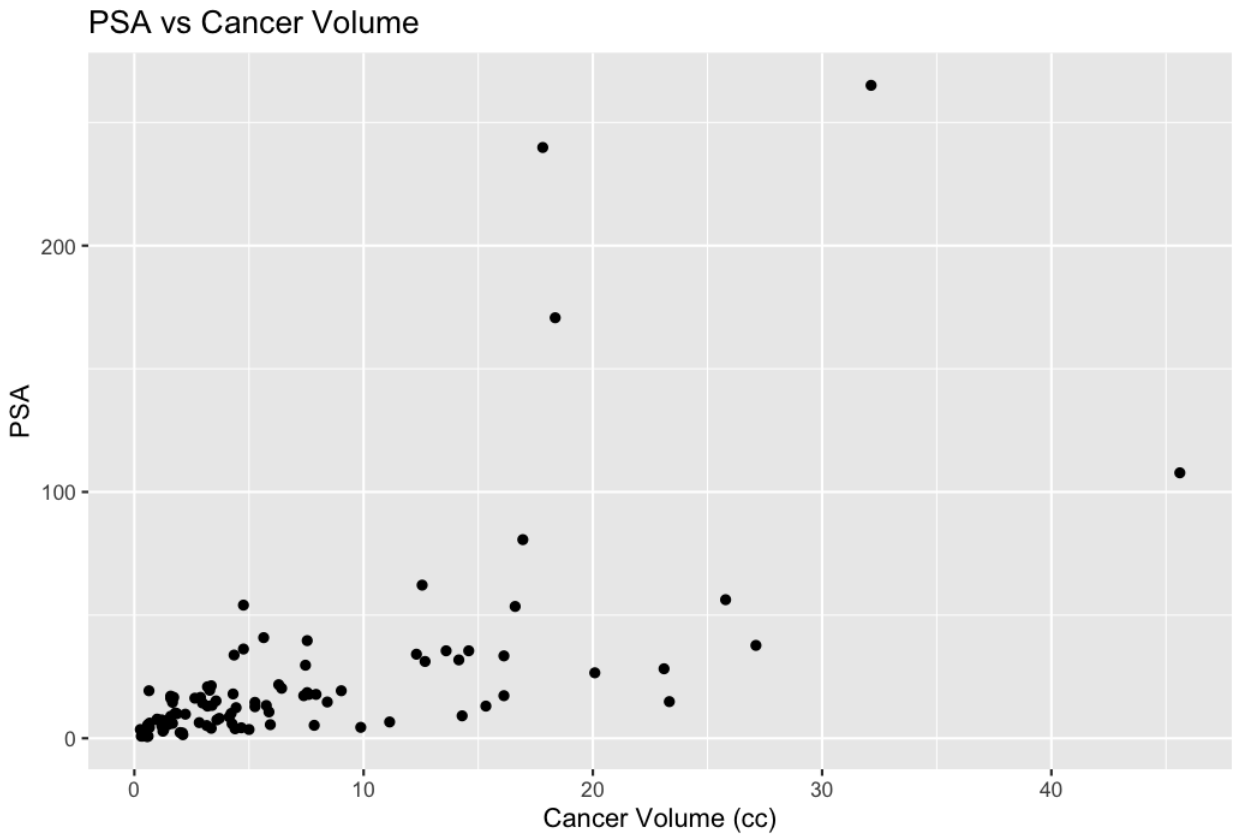

Figure 2: Cancer Volume vs PSA Level

Figure 2 shows prostate cancer volume in relation to PSA. There appears to be a moderately strong, positive, linear relationship between cancer volume and PSA (correlation coefficients: 0.62). This relationship is an indicator that cancer volume will likely be a useful predictor in determining PSA values. Additionally, we can see the three aforementioned extreme PSA values sticking above the rest of the y-axis, and a high reading for cancer volume sticking out on the x axis. This high cancer reading has the fourth highest PSA reading and looks to be following the linear trend seen with the rest of the data.

PSA vs Weight

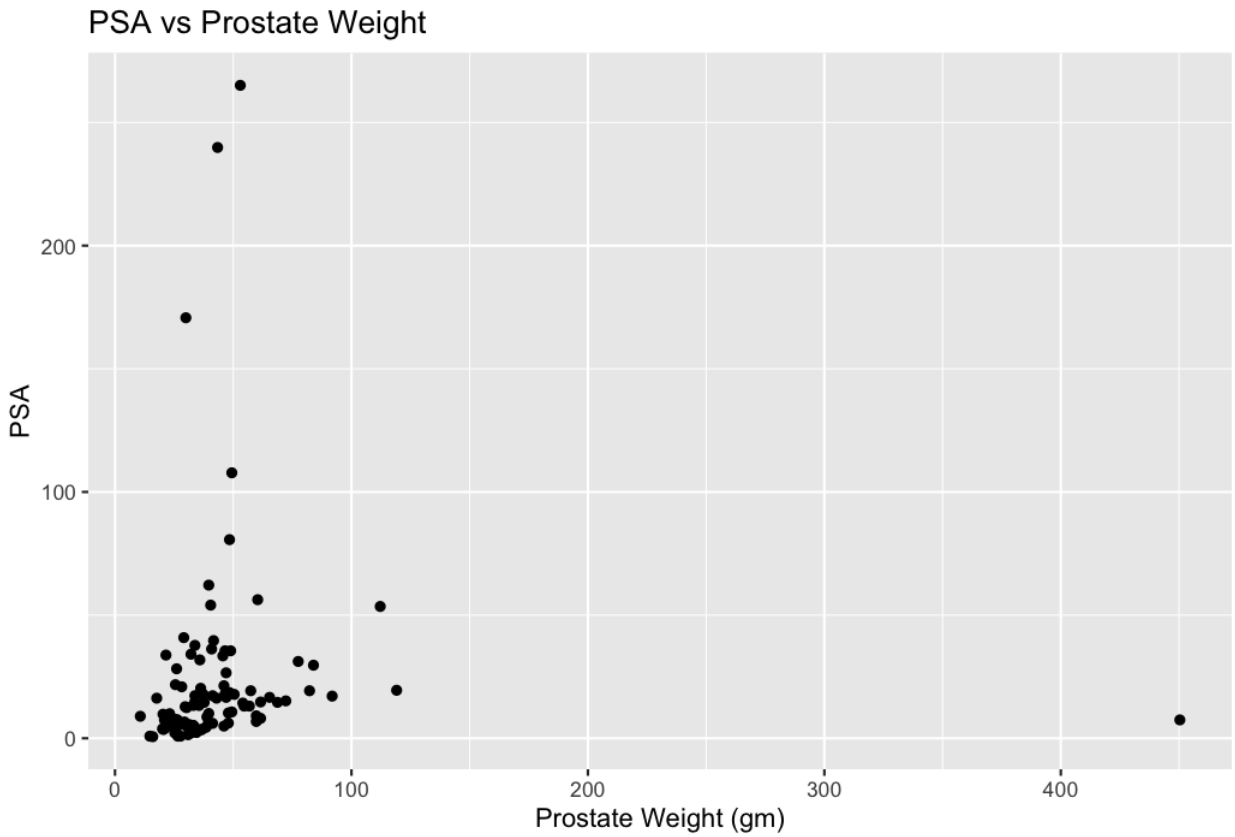

Figure 3: Weight vs PSA Level

Figure 3 shows the relationship between prostate weight and PSA levels. There appears to be a moderate positive linear relationship between prostate weight and PSA, suggesting that weight may be a reasonable predictor of PSA. However, there is one extreme outlier in terms of weight, around 450 grams—approximately four times higher than the next highest reading. Once again, the three extreme PSA readings deviate from the overall trend observed in the other data points. These outliers likely distort the correlation coefficient, making the relationship appear weaker than it is. The correlation coefficient, including all values, is r = 0.03, suggesting no obvious relationship. However, after removing the weight outlier, the r value increases to 0.15, indicating a slightly stronger but still somewhat weak relationship.

PSA vs Age

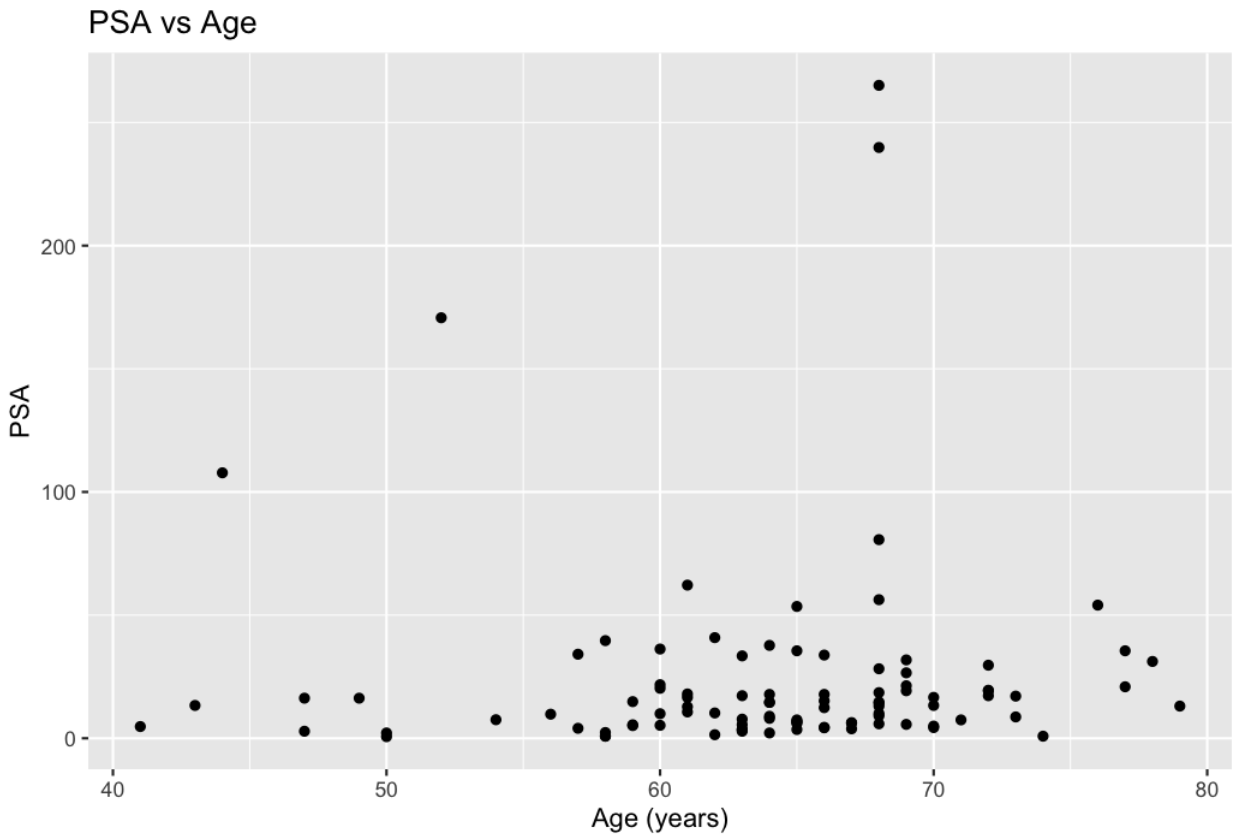

Figure 4: Age vs PSA Level

Figure 4 shows the relationship between age and PSA levels. There does not appear to be a linear relationship, as PSA readings remain similar across different ages. This is supported by the correlation coefficient of r = 0.02, suggesting that age is unlikely to be a significant predictor of PSA. Additionally, the density of points on the graph indicates that most subjects in the dataset are between 60 and 70 years old, although a histogram would offer a more appropriate representation of this age distribution.

PSA vs Hyperplasia

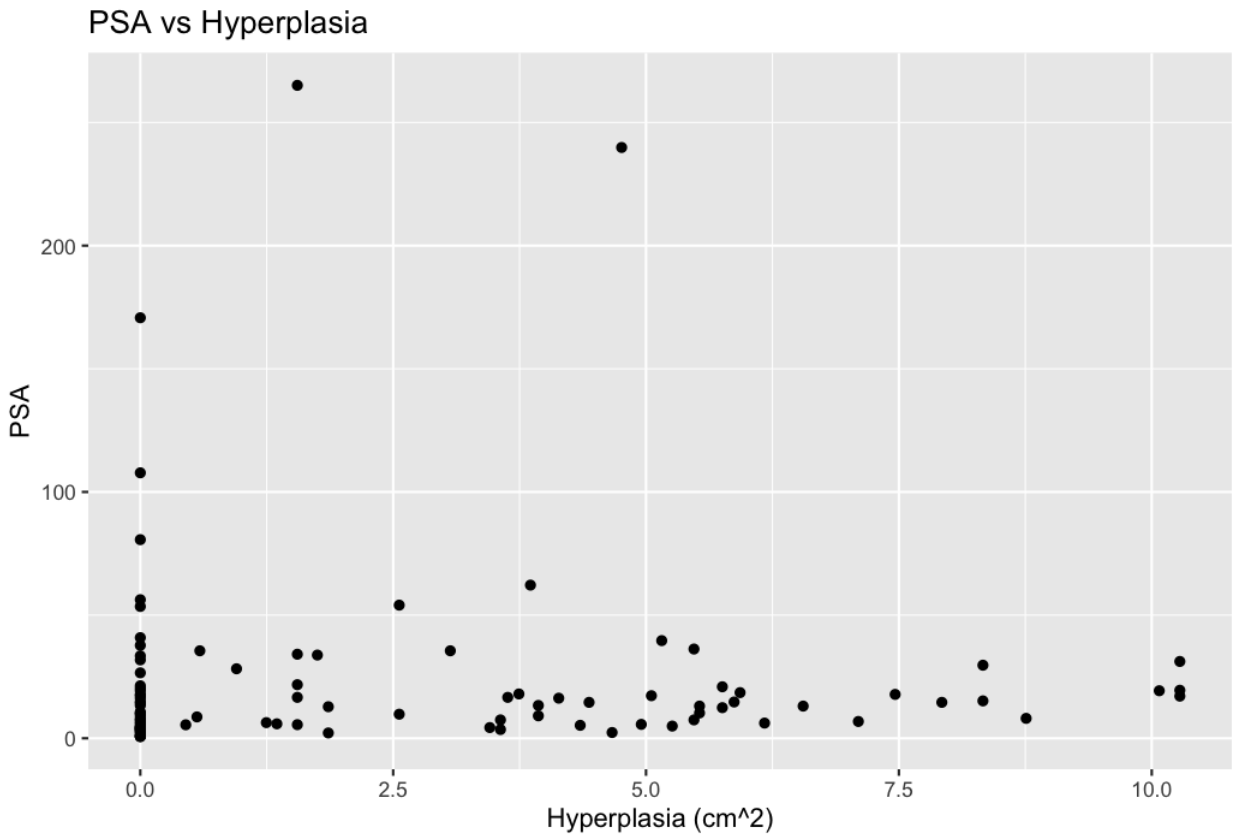

Figure 5: Hyperplasia vs PSA Level

Figure 5 shows the relationship between the amount of benign prostatic hyperplasia and PSA levels. Visually, there appears to be a bordering negligible, negative linear relationship ($r = -0.02$) since PSA readings are generally consistent across different hyperplasia readings. Additionally, there are many hyperplasia readings at 0, which also include a wide range of corresponding PSA values.

PSA vs Seminal

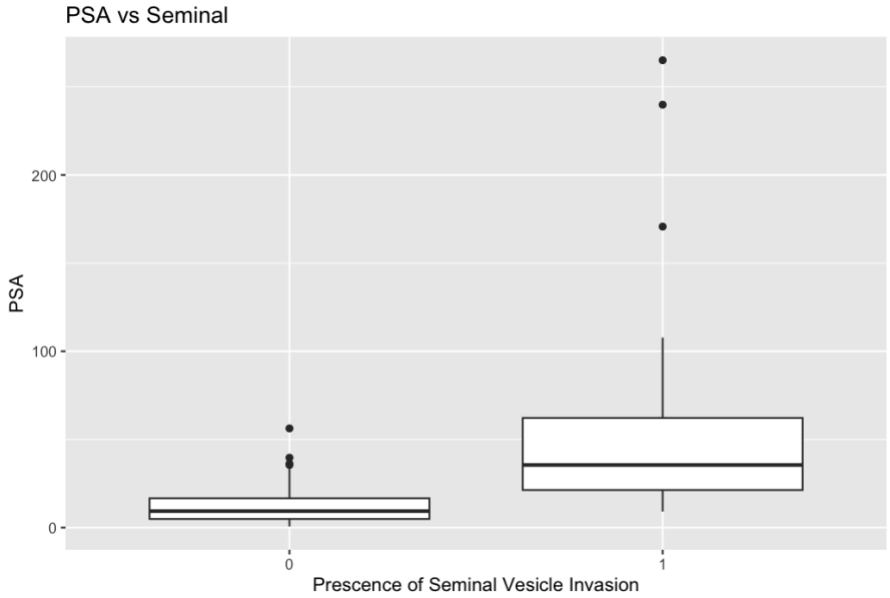

Figure 6: Seminal vs PSA Level

Figure 6 shows the distributions of PSA levels for both the presence (1) and absence (0) of seminal vesicle invasion. PSA levels are generally higher when a seminal vesicle is present (1) compared to when it is absent (0). Notably, the lower quartile of PSA levels when a seminal vesicle is present is higher than the upper quartile when it is absent. Based on this, a t-test was conducted with the null hypothesis stating that there is no difference in the mean PSA levels between the two groups (seminal = 1 and seminal = 0), and the alternative hypothesis stating that the mean PSA for the group with seminal = 1 is greater than the mean PSA for the group with seminal = 0. The t-test yielded a p-value of 0.001924. Since the p-value is less than 0.05, the null hypothesis is rejected, meaning that the mean PSA levels for the group with a seminal vesicle invasion are significantly greater than that for the group without one. Therefore, the presence or absence of a seminal vesicle invasion is likely a significant predictor of PSA levels in a model.

```
        Welch Two Sample t-test

data:  psa by seminal
t = 3.263, df = 20.242, p-value = 0.001924
alternative hypothesis: true difference in means between group 1 and group 0 is greater than 0
95 percent confidence interval:
 24.5652     Inf
sample estimates:
mean in group 1 mean in group 0
       64.53086        12.45625
```

Figure 7: T-Test (PSA Values for Seminal Groups)

PSA vs Capsular

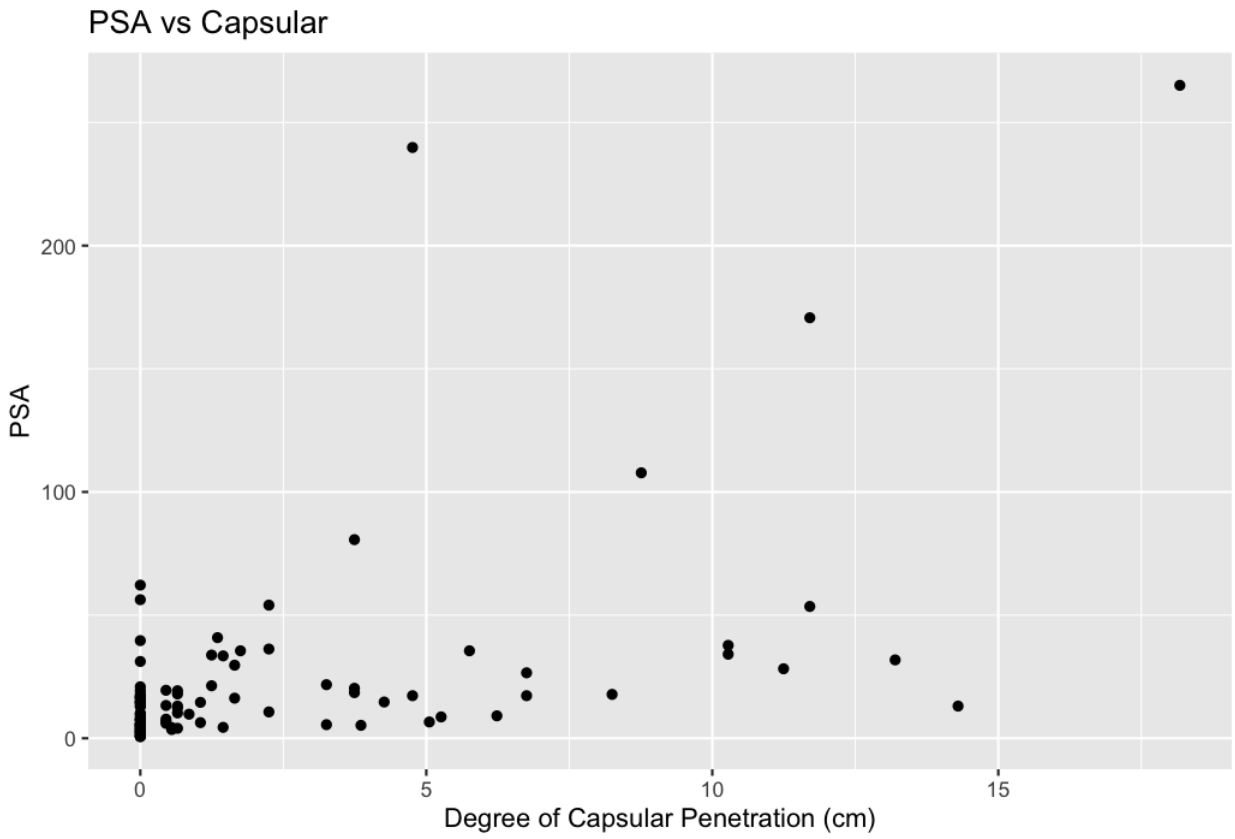

Figure 8: Capsular vs PSA Level

Figure 8 shows the relationship between the degree of capsular penetration and PSA levels. There is a moderately strong, positive linear relationship between the two variables, supported by a correlation coefficient of r = 0.55, which suggests that capsular penetration may be a good predictor of PSA levels. Additionally, a large number of observations have a value of 0 for capsular penetration, indicating many subjects did not have any degree of capsular penetration.

PSA vs Score

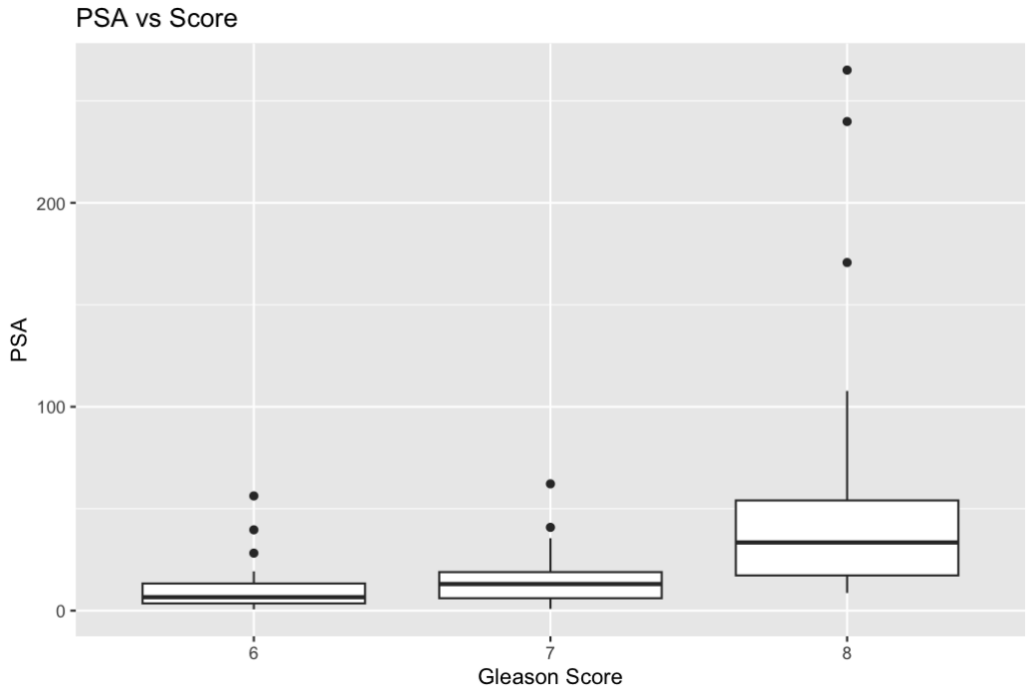

Figure 9: Gleason Score vs PSA Level

Figure 9 shows the distributions of PSA levels for various Gleason scores (with higher scores indicating a worse prognosis). PSA levels tend to increase as Gleason scores increase. Specifically, the lower quartile of PSA values for a Gleason score of 8 is above or near the upper quartile for Gleason scores of 6 and 7. While Gleason scores of 7 have a higher median PSA value than 6, the difference is not large. Given the differing PSA distributions across Gleason scores, an ANOVA test is appropriate (H_0: the mean PSA levels are the same across all levels of Gleason score; H_1: at least one group has a mean PSA level different from the others). The ANOVA test returns a p-value of 1.13e-05. Since the p-value is less than 0.05, the null hypothesis is rejected, indicating that there is a statistically significant difference in the mean PSA levels across the different Gleason score groups. Although the ANOVA indicates a significant difference in PSA readings across the Gleason scores, a t-test would be useful to examine the differences between specific Gleason scores. Nevertheless, the ANOVA results suggest that Gleason score is likely a strong predictor of PSA levels.

```
            Df Sum Sq Mean Sq F value   Pr(>F)
score        1  29466   29466    21.5 1.13e-05 ***
Residuals   95 130206    1371
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 10: ANOVA (PSA Values for Gleason Score Groups)

Pair Plot

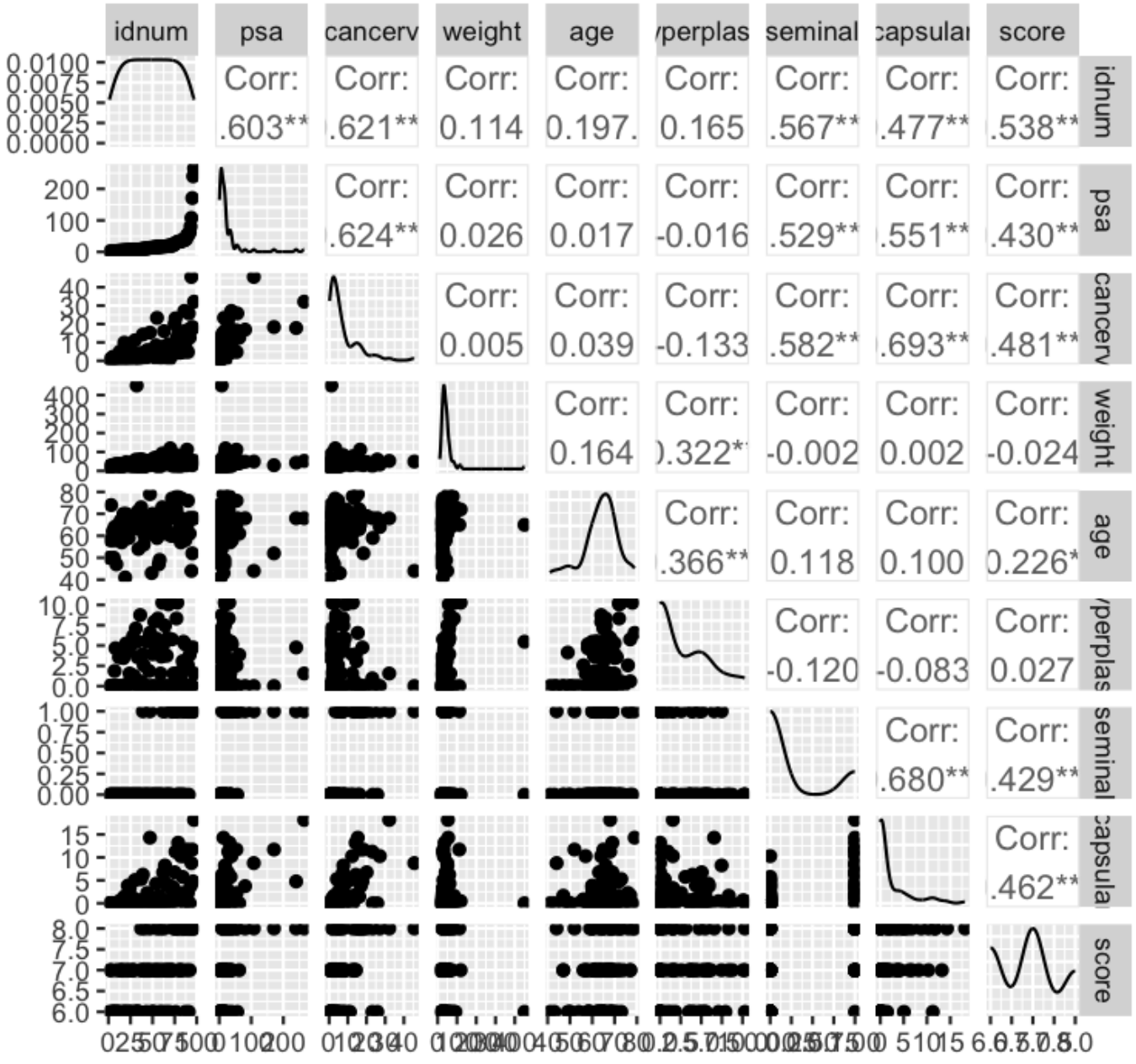

Figure 11: Pair Plot

Figure 11 is a pair plot, which is useful for visualizing correlations and distributions and examining potential relationships/multicollinearity between variables. In the cancer dataset, there does not appear to be significant multicollinearity, although seminal and capsular show a moderate correlation of 0.68.

Preprocessing

After conducting exploratory and basic statistical analysis, the next step is to use the variables to predict PSA. However, before building any predictive models, there are a few considerations. First, as mentioned earlier, there were no NA or NULL values in the dataset, so no action is required there. Additionally, there was no immediate need for feature engineering, so no new features were created. As observed in the graphs, however, there are numerous outliers. I chose to fit a model that includes the outliers initially and assess their impact on the model, rather than removing them beforehand. Lastly, the categorical variables *seminal* and *score* were converted to factors, and the *idnum* column was removed, as it is a identifier variable should not influence the dependent variable.

Modeling

Saturated Model 1

A saturated model, or a model using all variables, was used as a baseline model. The r output is as follows:

```
Call:
lm(formula = psa ~ ., data = cancer1)

Residuals:
    Min      1Q  Median      3Q     Max
-68.153  -7.323  -0.177   6.403 161.547

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 31.849265  28.958981   1.100  0.27442
cancerv      1.748107   0.615858   2.838  0.00563 **
weight      -0.004546   0.074038  -0.061  0.95118
age         -0.537278   0.471991  -1.138  0.25808
hyperplasia  1.530782   1.201007   1.275  0.20581
seminal1    21.108723  10.844893   1.946  0.05479 .
capsular     1.097882   1.322879   0.830  0.40883
score7      -1.661862   7.570741  -0.220  0.82676
score8      18.423157  10.661795   1.728  0.08750 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30.91 on 88 degrees of freedom
Multiple R-squared:  0.4733,    Adjusted R-squared:  0.4254
F-statistic: 9.886 on 8 and 88 DF,  p-value: 1.037e-09
```

Figure 12: Saturated Model 1

The equation for the saturated model:

*PSA = 31.85 + 1.75 * cancerv - 0.0045 * weight - 0.537 * age + 1.53 * hyperplasia + 21.11 * seminal + 1.10 * capsular - 1.66 * score7 + 18.42 * score8*

There are a few notable issues with the saturated model. First, there were only three significance predictors (*cancerv, seminal, & score8*) of PSA at a 0.10 significance level, and only one (*cancerv*) at a 0.05 significance level. This suggests that many variables in the model may not be contributing meaningfully to the prediction of PSA. Additionally, the adjusted R-squared value is 0.4254, indicating that only 42.54% of the variation in PSA levels can be explained by the model. While this is a relatively low explanatory power, the model is still meaningful, as the F-statistic has a p-value of 1.037e-09, significantly less than 0.05, suggesting that the model is

significantly better than a model with no predictors. Beyond the coefficients and their significance, however, it is also important to evaluate other parts of the model.

Variable	GVIF	df	GVIF^(1/(2*df))
cancerv	2.366412	1	1.538315
weight	1.150308	1	1.072524
age	1.240489	1	1.113773
hyperplasia	1.331359	1	1.153845
seminal	2.024987	1	1.423020
capsular	2.516346	1	1.586299
score	1.708552	2	1.143292

Table 3: VIF - Saturated Model 1

The VIF table shows the degree of multicollinearity among the predictor variables in the regression model. Most variables, such as *weight*, *age*, and *hyperplasia*, have low VIFs (less than 2), suggesting minimal collinearity with other predictors. The variables *cancerv*, *seminal*, and *capsular* have moderately low VIFs (ranging from 2.02 to 2.52), indicating some collinearity, but not to a concerning degree. Overall, there is not severe multicollinearity in the model, and none of the VIFs exceed the typical threshold of 5 or 10 that would warrant further investigation or variable removal.

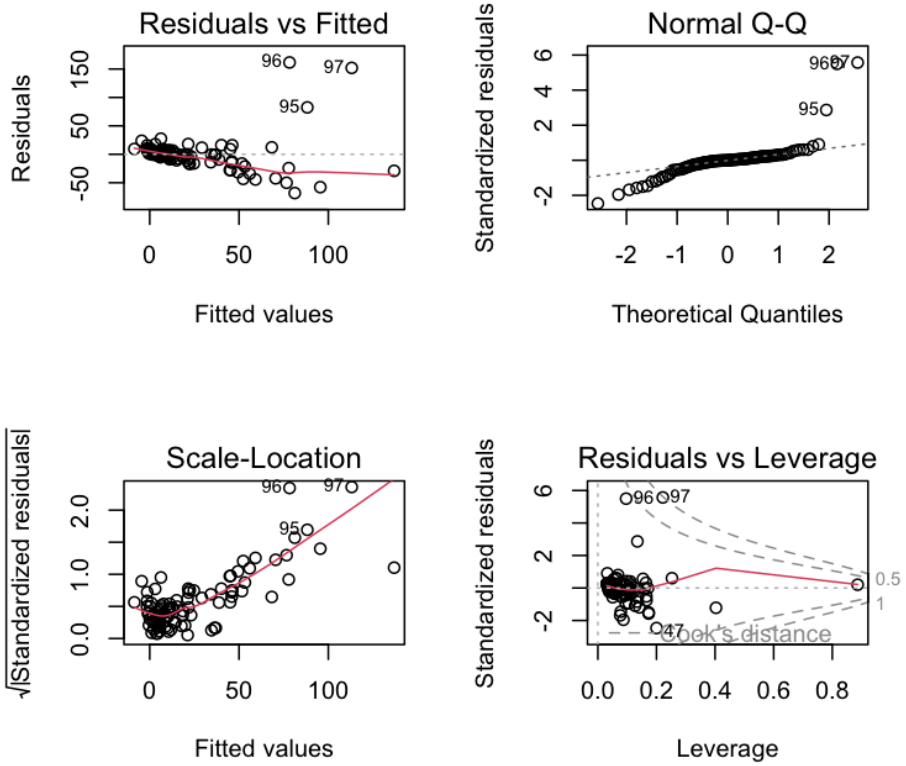

Figure 13: Diagnostic Plots - Saturated Model 1

Based on the diagnostic plots, several issues are affecting the model's performance. In the residual vs. fitted plot, a potential pattern in the residuals is visible, though it does not appear to be particularly extreme and concerning. However, the most notable observations in the plots are points 95, 96, and 97—the individuals with extremely high PSA levels. These points have PSA levels of 171, 240, and 265 mg/ml, respectively, whereas the mean PSA level for the rest of the dataset is 17.30 mg/ml. Surprisingly, point 97 is the only one with a Cook's Distance greater than 0.5, at 0.98. For Studentized Residuals, only points 96 and 97 are greater than the absolute value of 3, coming in at 6.75 and 6.89, respectively. Given the limited number of extreme values, I am hesitant to remove data. However, since the goal is to develop the most accurate model for the typical patient with advanced prostate cancer, I removed points 96 and 97 and re-fit the saturated model.

Saturated Model 2

After removing points 96 and 97, the saturated model was refit. The r output is as follows:

```
Call:
lm(formula = psa ~ ., data = cancer2)

Residuals:
    Min      1Q  Median      3Q     Max
-35.240  -6.182  -0.038   3.576 110.786

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 40.52074   15.80239   2.564 0.012077 *
cancerv      1.33287    0.33680   3.957 0.000155 ***
weight       0.01378    0.04039   0.341 0.733832
age         -0.62216    0.25737  -2.417 0.017746 *
hyperplasia  0.81135    0.65914   1.231 0.221703
seminal1    19.18263    6.08573   3.152 0.002232 **
capsular    -0.60131    0.77264  -0.778 0.438552
score7       2.27165    4.13311   0.550 0.584002
score8      14.73705    5.84752   2.520 0.013575 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.84 on 86 degrees of freedom
Multiple R-squared:  0.5356,    Adjusted R-squared:  0.4924
F-statistic:  12.4 on 8 and 86 DF,  p-value: 1.173e-11
```

Figure 14: Saturated Model 2

The equation for the saturated model:

*PSA = 40.52 + 1.33 * cancerv + 0.0138 * weight - 0.622 * age + 0.811 * hyperplasia + 19.18 * seminal1 - 0.601 * capsular + 2.27 * score7 + 14.74 * score8*

This time, more coefficients are significant in the model. First, the intercept is significant. As with the previous model, *cancerv*, *seminal*, and *score8* are significant at the 0.05 significance level. *Score7*, as in the previous model, is not significantly different from the baseline category, *score6*. However, since *score8* is significant, it is important to retain the entire variable in the model. Additionally, *age* is significant at the 0.05 significance level as well. In terms of model performance, the adjusted R^2 increased to 0.4924, indicating that approximately 7% more of the variation in PSA levels is now explained by the model.

Variable	GVIF	df	GVIF^(1/(2*df))
cancerv	2.080052	1	1.442239
weight	1.153440	1	1.073983
age	1.234834	1	1.111231
hyperplasia	1.342292	1	1.158573
seminal	1.985182	1	1.408965
capsular	2.337721	1	1.528961
score	1.618936	2	1.127996

Table 4: VIF - Saturated Model 2

From the first saturated model, the VIF values have decreased, indicating a reduced degree of multicollinearity between the variables and suggesting each variable contributes more independently to the model, with less overlap in the information they provide.

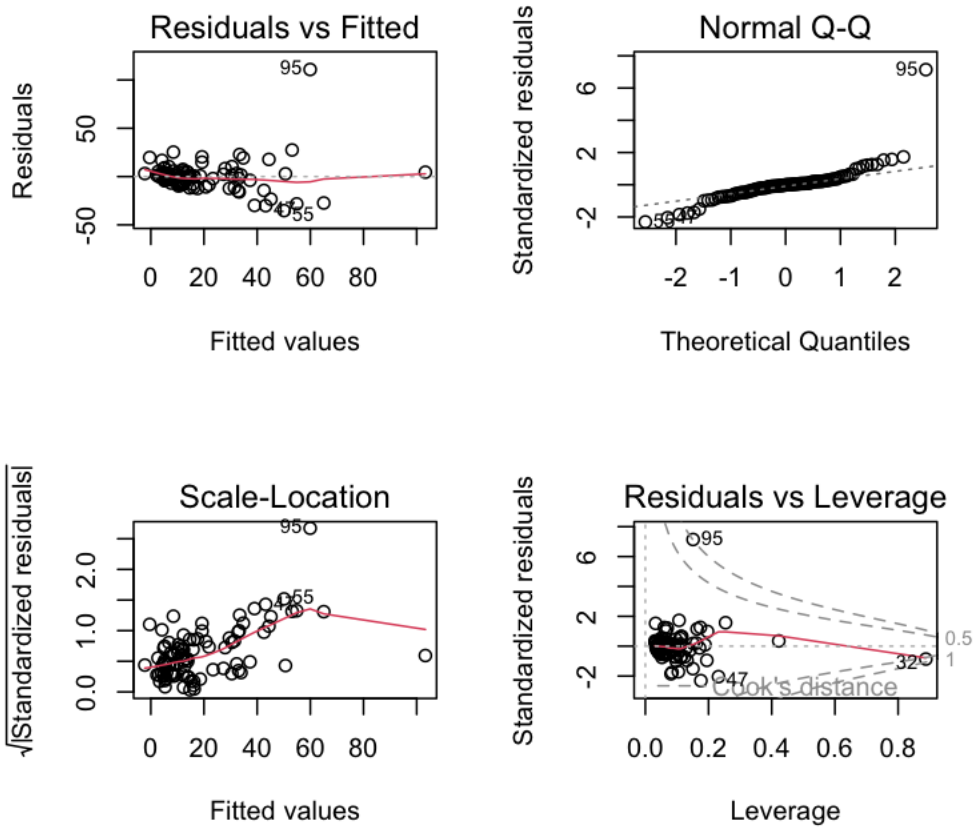

Figure 15: Diagnostic Plots - Saturated Model 2

There is no discernible pattern in the residuals vs. fitted plot, indicating that heteroscedasticity is not a concern this time. Additionally, the Q-Q plot suggests that the residuals are approximately normally distributed. However, there are outliers at points 95 and potentially 32. Examining the Cook's Distance for these points, both are above 0.5, with point 32 having a value of 0.65 and point 95 having a value of 1.00. Additionally, the Studentized Residual for point 95 is 11.13, well beyond the threshold for an outlier. In contrast, point 32's Studentized Residual is -0.86, well below the threshold for outliers. Point 32 corresponds to a subject with a prostate weight of 450 grams, over 10 times the median prostate weight and nearly four times larger than the next largest prostate weight in the dataset. This extreme value could either be due to measurement error or represent an unusually large prostate. Regardless, this point is not representative of the rest of the data or typical prostate size even for individuals with advanced prostate cancer, and since the goal is to create the most robust model for the average person with advanced prostate cancer, both points 32 and point 95 were removed from the dataset.

Saturated Model 3

After removing points 32 and 95, the saturated model was refit. The r output is as follows:

```
Call:
lm(formula = psa ~ ., data = cancer3)

Residuals:
    Min      1Q  Median      3Q     Max
-33.161  -4.336  -1.443   2.715  31.887

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.70990   10.18083   1.543 0.126569
cancerv      1.51033    0.21551   7.008 5.61e-10 ***
weight       0.16267    0.07438   2.187 0.031514 *
age         -0.28766    0.16714  -1.721 0.088912 .
hyperplasia  0.23147    0.47507   0.487 0.627359
seminal1    17.04174    3.83944   4.439 2.73e-05 ***
capsular    -1.72847    0.49704  -3.478 0.000804 ***
score7       2.88803    2.60311   1.109 0.270401
score8       8.56235    3.74098   2.289 0.024601 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.6 on 84 degrees of freedom
Multiple R-squared:  0.6758,    Adjusted R-squared:  0.6449
F-statistic: 21.89 on 8 and 84 DF,  p-value: < 2.2e-16
```

Figure 16: Saturated Model 3

The equation for the saturated model:

*PSA = 15.71 + 1.51 * cancerv + 0.16 * weight - 0.29 * age + 0.23 * hyperplasia + 17.04 * seminal1 - 1.73 * capsular + 2.89 * score7 + 8.56 * score8*

In the third saturated model, which excludes points 32, 95, 96, and 97, all variables except *hyperplasia* are significant at least at the 0.10 significance level, and all but *age* are significant at the 0.05 significance level. As in the previous models, *score7* remains not significant, while *score8* is significant. However, unlike the previous model, the intercept is not statistically significant. Regardless, the adjusted R^2 increased to 0.6449, an approximately 15% increase in the amount of variability in PSA levels explained by the model. Additionally, as with the other model, no variables had a high VIF.

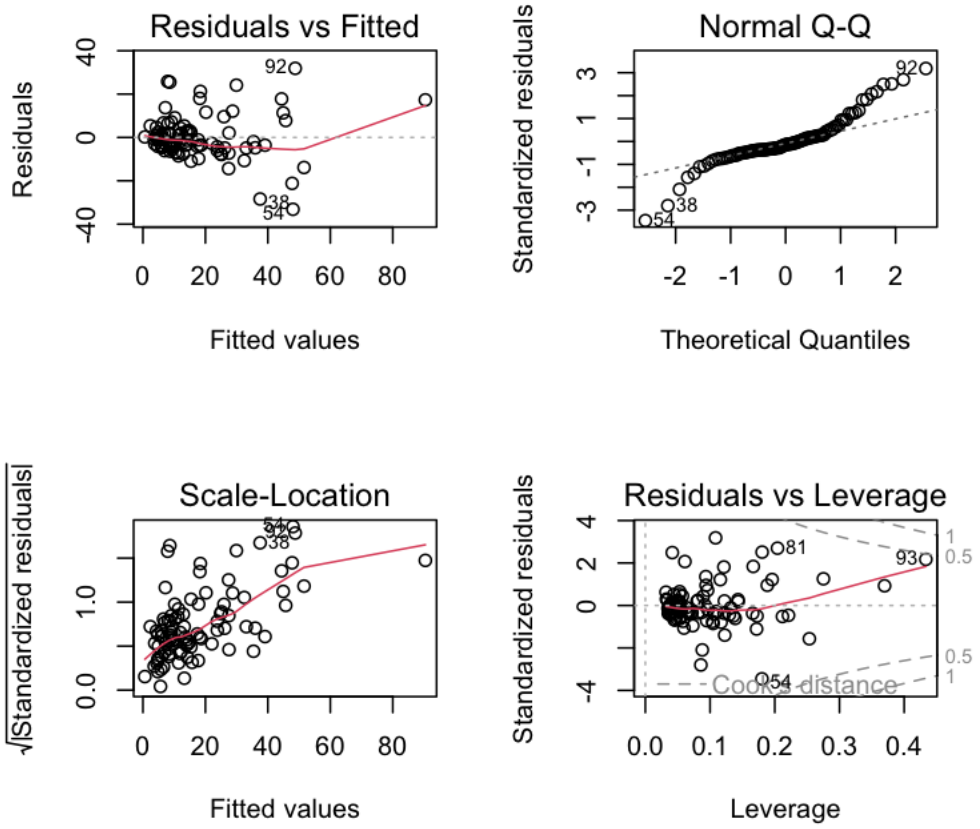

Figure 17: Diagnostic Plots - Saturated Model 3

There is no discernible pattern in the residuals vs. fitted values plot, and the residuals appear to be normally distributed, as seen in the Q-Q plot. Additionally, there are no obvious outliers. None of the points have a Cook's Distance greater than 0.5. While points 54 and 92 have Studentized Residuals with absolute values greater than 3, they are not sufficiently different from the rest of the data to warrant removal, especially considering the already small dataset.

Data Point	PSA	Cancerv	Weight	Age	Hyperplasia	Seminal	Capsular	Score
54	14.9	23.3	33.8	59	0	0	0	8
92	80.6	16.9	48.4	68	0	1	3.74	8
Median (excluding points 54, 92)	36.6	65.0	1.35	0	1.35	0.4493	-	-

Table 5: Potential Outliers – Saturated Model 3

Native Model 1

Saturated models are often overly complex and include variables that are not statistically significant. By removing these non-significant variables, the model typically either performs better or maintain similar performance while reducing complexity. To select variables, I am first applying a straightforward approach: using only the variables that were statistically significant in the final saturated model.

```
Call:
lm(formula = psa ~ cancerv + weight + age + seminal + capsular +
    score, data = cancer3)

Residuals:
    Min      1Q  Median      3Q     Max
-33.181  -4.235  -1.860   3.189  31.571

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.4043     9.7777   1.473 0.144397
cancerv       1.4980     0.2131   7.031 4.83e-10 ***
weight        0.1825     0.0619   2.949 0.004116 **
age          -0.2700     0.1624  -1.662 0.100147
seminal1     16.7716     3.7821   4.434 2.74e-05 ***
capsular     -1.7187     0.4944  -3.476 0.000803 ***
score7        3.0417     2.5723   1.182 0.240305
score8        8.4604     3.7183   2.275 0.025403 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.56 on 85 degrees of freedom
Multiple R-squared:  0.6749,    Adjusted R-squared:  0.6481
F-statistic:  25.2 on 7 and 85 DF,  p-value: < 2.2e-16
```

Figure 18: Naïve Model 1

The equation for the naive model:

*PSA = 14.4043 + 1.4980 * cancerv + 0.1825 * weight - 0.2700 * age + 16.7716 * seminal1 - 1.7187 * capsular + 3.0417 * score7 + 8.4604 * score8*

All significant variables (at the 0.10 significance level) from the final saturated model were included in the first simplified model. *Age*, which was significant at the 0.10 level in the final saturated model, has a p-value of 0.1001, just above the 0.10 threshold. Additionally, similar to the final saturated model, the intercept is not statistically significant. The adjusted R^2 is slightly higher, 0.6481, compared to the final saturated model's value of 0.6449. While there is only a modest improvement in performance, the model is now simpler, with no trade-off in performance due to the reduced complexity. Like the final saturated model, no variables had a VIF greater than 2.5 nor were there any issues with the diagnostic plots –no overly concerning outliers, and the residuals are normally distributed and random.

Naïve Model 2

The second naïve model was fit with all the variables as the first naïve model except *age*, which was not significant in the first naïve model.

```
Call:
lm(formula = psa ~ cancerv + weight + seminal + capsular + score,
    data = cancer3)

Residuals:
    Min      1Q  Median      3Q     Max
-32.528  -4.973  -2.000   3.607  31.154

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.09120    2.98030  -0.366 0.715161
cancerv      1.54698    0.21316   7.257 1.64e-10 ***
weight       0.14873    0.05906   2.518 0.013643 *
seminal1    16.38693    3.81354   4.297 4.54e-05 ***
capsular    -1.76308    0.49871  -3.535 0.000659 ***
score7       2.37808    2.56705   0.926 0.356838
score8       7.37383    3.69775   1.994 0.049305 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.66 on 86 degrees of freedom
Multiple R-squared:  0.6643,    Adjusted R-squared:  0.6409
F-statistic: 28.36 on 6 and 86 DF,  p-value: < 2.2e-16
```

Figure 19: Naïve Model 2

The equation for the naive model:

*PSA = -1.09120 + 1.54698 * cancerv + 0.14873 * weight + 16.38693 * seminal1 - 1.76308 * capsular + 2.37808 * score7 + 7.37383 * score8*

All coefficients in the model are statistically significant, except for *score7*. As with most of the models, the intercept is also not significant. The Variance Inflation Factor (VIF) values for all variables are below 2.5, indicating no issues with multicollinearity. Additionally, there are no extreme outliers, and the residuals appear to be normally distributed and randomly scattered. However, the adjusted R^2 of 0.6409 is slightly lower than that of the first naïve model, which had an adjusted R^2 of 0.6481. To determine if the added complexity of including *age* in the model is justified by an improvement in performance, an ANOVA test can be used.

```
Analysis of Variance Table

Model 1: psa ~ cancerv + weight + seminal + capsular + score
Model 2: psa ~ cancerv + weight + age + seminal + capsular + score
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     86 9778.9
2     85 9471.0  1    307.87 2.7631 0.1001
```

Figure 20: ANOVA Comparing Naïve Model 1 & 2

Since the p-value for *age* (0.1001) is greater than 0.05, the null hypothesis cannot be rejected. This suggests that adding the *age* variable does not significantly improve the model. In other words, Naïve Model 2 or the simpler model, which includes only *cancerv*, *weight*, *seminal*, *capsular*, and *score*, is not significantly worse than the more complex model that includes the additional *age* variable.

Ridge Regression

While the naïve models performed well, naively selecting coefficients is often time-consuming and error prone due to human judgment. In linear regression, there are several notable methods for feature selection, including forward and backward stepwise selection using AIC, BIC, adjusted R^2, or Mallows' Cp. However, in this case, ridge and lasso regression will be used.

Model Performance Metrics:

- **R-squared:** 0.6696
- **Adjusted R-squared:** 0.6381
- **Mean Squared Error (MSE):** 149.261
- **Root Mean Squared Error (RMSE):** 12.2172
- **Optimal Lambda (λ):** 1.2904

Variable	Coefficient Estimate
Intercept	15.0367
Cancerv	1.3237
Weight	0.1581
Age	-0.2565
Hyperplasia	0.1655
Seminal1	14.9783
Capsular	-1.2133
Score7	2.2380
Score8	8.2644

Table 6: Ridge Regression Coefficients

The equation for the ridge regression model:

*PSA = 15.0367 + 1.3237 * cancerv + 0.1581 * weight - 0.2565 * age + 0.1655 * hyperplasia + 14.9783 * seminal1 - 1.2133 * capsular + 2.2380 * score7 + 8.2644 * score8*

The ridge regression model performs slightly worse than both the best naïve model and the saturated model in terms of adjusted R^2 (0.6381). However, the majority of the coefficients are very similar to those from the final saturated and naïve models, with *seminal* and *score8* having the largest coefficients.

Lasso Regression

Model Performance Metrics:

- **R-squared:** 0.6758
- **Adjusted R-squared:** 0.6449
- **Mean Squared Error (MSE):** 149.6969
- **Root Mean Squared Error (RMSE):** 12.2351
- **Optimal Lambda (λ):** 0.0231

Variable	Coefficient Estimate
Intercept	15.4146
Cancerv	1.5043
Weight	0.1621
Age	-0.2809
Hyperplasia	0.2196
Seminal1	16.8973
Capsular	-1.6979
Score7	2.7849
Score8	8.4287

Table 7: Lasso Regression Coefficients

The equation for the lasso regression model:

*PSA = 15.4146 + 1.5043 * cancerv + 0.1621 * weight - 0.2809 * age + 0.2196 * hyperplasia + 16.8973 * seminal1 - 1.6979 * capsular + 2.7849 * score7 + 8.4287 * score8*

The lasso regression model performs very similar to the ridge regression and final saturated and naïve models with an adjusted R^2 of .6449. Surprisingly, *hyperplasia* -- which was not a significant variable in all other models -- was not set to zero by the L1 regularization. In fact, all coefficients were similar to the ridge regression model and other models, too.

Model Comparison

The first and second saturated models that were trained with the outliers performed significantly worse than the other models in terms of adjusted R^2 and therefore will be left out of the comparison. Additionally, since they have additional observations, it would not be fair to compare their RMSE's to other models that have fewer observations. After removing the outliers, all models performed similarly across various metrics. However, it is important to note that R^2 is not always a reliable comparison, as the models contain different numbers of predictor variables. Additionally, each model was also evaluated using 10-fold cross-validation.

Model	R^2	Adjusted R^2	R^2 (10-fold CV)	RMSE (10-fold CV)
Saturated Model 3	0.6758	0.6449	0.6009	11.62893
Naive Model 1	0.6749	0.6481	0.5608	11.73791
Naive Model 2	0.6643	0.6409	0.5914	11.62934
Ridge Regression	0.6696	0.6381	0.6641	10.8049
Lasso Regression	0.6758	0.6449	0.5865	11.40518

Table 8: Model Comparison

The ridge regression model performs similarly to other models in terms of R^2 and adjusted R^2 but stands out with the lowest RMSE in 10-fold cross-validation, indicating it generalizes the best to unseen data. It outperforms Naive Models 1 and 2, which show signs of minor overfitting, as evidenced by their higher cross validation RMSE. Saturated Model 3 and the lasso regression model are comparable to the ridge regression model in terms of R^2 (same number of variables), but the lasso regression model has a slightly higher RMSE, suggesting it doesn't generalize as well as the ridge regression model. Overall, the ridge regression model offers the best balance of model fit and generalization, while other models like the saturated model and lasso regression perform similarly but with slightly higher error on unseen data.

Conclusions

Through both the graphs and the linear regression model, PSA levels were able to be partially explained by other variables. However, the influence of these variables on PSA levels varied. Notably, the amount of benign prostatic hyperplasia appeared to have little to no impact on PSA levels, while all other variables had at least a slight influence. Across all models and shown by the graphs, the presence or absence of a seminal vesicle invasion, whether a subject had a Gleason score of 8, and cancer volume were consistently significant predictors of PSA levels. Other variables, such as prostate weight, age, and capsular penetration played smaller roles in predicting PSA levels, too.

To predict PSA levels, the ridge regression model provides the best balance of fit and generalization and would be the best model for predicting on new data. The interpretation of the ridge regression model is as follows:

*PSA = 15.0367 + 1.3237 * cancerv + 0.1581 * weight - 0.2565 * age + 0.1655 * hyperplasia + 14.9783 * seminal1 - 1.2133 * capsular + 2.2380 * score7 + 8.2644 * score8*

- *Intercept*: When all continuous variables are 0 and the categorical variables are at their reference category (seminal = 0, score = 6), the predicted PSA level will be **15.0367** mg/ml.
- *Cancerv*: Holding all other variables constant, for each additional cc of prostate cancer volume, PSA levels will increase by **1.3237** mg/ml on average.
- *Weight*: Holding all other variables constant, for each additional gram increase in prostate weight, PSA levels will increase by **0.1581** mg/ml on average.
- *Age*: Holding all other variables constant, for each additional year of age, PSA levels will decrease by **0.2565** mg/ml on average.
- *Hyperplasia*: Holding all other variables constant, for each additional cm^2 of benign prostatic hyperplasia, PSA levels will decrease by **0.1655** mg/ml on average.
- *Seminal1*: Holding all other variables constant, if the subject has seminal vesicle invasion, the PSA level will increase by **14.9783** mg/ml on average compared to subjects without seminal vesicle invasion.
- *Capsular*: Holding all other variables constant, for each additional cm of capsular penetration, PSA levels will decrease by **1.2133** mg/ml on average.
- *Score7*: Holding all other variables constant, if the subject has a Gleason score of 7, the PSA level will increase by **2.2380** mg/ml on average compared to subjects with a Gleason score of 6.
- *Score8*: Holding all other variables constant, if the subject has a Gleason score of 8, the PSA level will increase by **8.2644** mg/ml on average compared to subjects with a Gleason score of 6.

Sources of Error & Next Steps

There were several limitations to this statistical analysis. First, the sample size was relatively small, with only 97 observations, four of which were removed. As is often the case with experimental data, a larger sample size would have provided a more robust estimate and a clearer understanding of the true population. Additionally, although scaling is not always necessary for linear regression, it could have improved model performance in this case. Finally, while a linear regression model was a suitable choice given the linear nature of the data, alternative machine learning models, such as random forests or boosting methods, could have potentially improve predictive performance. However, these methods would likely require additional observations, too.

References

Bobbitt, Zach. 2019. “How to Calculate Variance Inflation Factor (VIF) in R.” Statology. May 9, 2019. https://www.statology.org/variance-inflation-factor-r/.

Dutta, Sreejata. 2019. Review of *Predicting Electrical Power Output in a Combined Cycle Power Plant*.

Huang, Yibi. n.d. “STAT 224 Lecture 18 Ridge and Lasso Regressions.” https://www.stat.uchicago.edu/~yibi/teaching/stat224/L18.pdf.

“Understanding Lasso and Ridge Regression | R-Bloggers.” 2020. R Bloggers. June 16, 2020. https://www.r-bloggers.com/2020/06/understanding-lasso-and-ridge-regression/.

Appendix (R code)

```
library(tidyverse)
library(gvlma)
library(lmtest)
library(glmnet)
library(caTools)
library(caret)
library(MASS)
library(readxl)
library(knitr)
library(kableExtra)
library(gridExtra)
library(broom)
library(car)

#loading data
cancer <- read_excel('filepath.xlsx')

## Exploratory Analysis & Preprocessing
head(cancer)
dim(cancer) #97 by 9
summary(cancer)
sapply(cancer, class)
sum(is.na(cancer))
```

```
### Plots
GGally::ggpairs(cancer[,])

#PSA
```{r}
ggplot(cancer, aes(x = psa)) +
  geom_boxplot(fill = "white", color = "black", outlier.colour = "red",
outlier.size = 2) +
  labs(
    title = "Boxplot of PSA Levels",
    x = "Prostate-Specific Antigen (PSA)",
    y = "Value"
  ) +
  scale_x_continuous(breaks = seq(0, max(cancer$psa), by = 20)) +
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
    axis.title.x = element_text(size = 12),
    axis.title.y = element_text(size = 12),
    axis.text = element_text(size = 10),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank()
  )

hist(cancer$psa, breaks = 100)
```
```

```
#PSA vs Cancer Volume
ggplot(cancer, aes(x = cancerv, y = psa)) +
  geom_point() +
  labs(
    title = "PSA vs Cancer Volume",
    x = "Cancer Volume (cc)",
    y = "PSA"
  )

cor(cancer$psa, cancer$cancerv)

#PSA vs Prostate Weight
ggplot(cancer, aes(weight, psa))+
  geom_point()+
  labs(
    title = "PSA vs Prostate Weight",
    x = "Prostate Weight (gm)",
    y = "PSA"
  )

cor(cancer$psa, cancer$weight)
cor(cancer[-32,]$psa, cancer[-32,]$weight)

#PSA vs Age
ggplot(cancer, aes(age, psa))+
  geom_point()+
  labs(
    title = "PSA vs Age",
    x = "Age (years)",
    y = "PSA"
  )

cor(cancer$psa, cancer$age)

#PSA vs Hyperplasia
ggplot(cancer, aes(hyperplasia, psa))+
  geom_point()+
  labs(
    title = "PSA vs Hyperplasia",
    x = "Hyperplasia (cm^2)",
    y = "PSA"
  )

cor(cancer$psa, cancer$hyperplasia)

#PSA vs Seminal
ggplot(cancer, aes(as.factor(seminal), psa))+
  geom_boxplot()+
  labs(
    title = "PSA vs Seminal",
    x = "Prescence of Seminal Vesicle Invasion",
    y = "PSA"
  )
```

```
#T-test to determine if the mean psa values for seminal = 1 are
significantly greater than those for seminal = 0.
cancert <- cancer
cancert$seminal <- factor(cancert$seminal, levels = c(1, 0))

seminal_t.test <- t.test(psa ~ seminal, data = cancert, alternative =
"greater")
seminal_t_test_table <- tidy(seminal_t.test)
print(seminal_t.test)
print(seminal_t_test_table)

#PSA vs Capsular
ggplot(cancer, aes(capsular, psa))+
  geom_point()+
  labs(
    title = "PSA vs Capsular",
    x = "Degree of Capsular Penetration (cm)",
    y = "PSA"
  )

cor(cancer$psa, cancer$capsular)

#PSA vs Gleason Score
ggplot(cancer, aes(as.factor(score), psa))+
  geom_boxplot()+
  labs(
    title = "PSA vs Score",
    x = "Gleason Score",
    y = "PSA"
  )

#Anova to determine if the mean psa values for each score are the same.
anova_score <- aov(psa ~ score, data = cancer)
summary(anova_score)
```

```
### Data Cleaning
#convert seminal to a factor
cancer$seminal = as.factor(cancer$seminal)

#convert score to a factor
cancer$score = as.factor(cancer$score)

#duplicate data set, remove idnum column
cancer1 <- cancer
cancer1$idnum <- NULL
#head(cancer1)

## Modeling

### Saturated Model
#all variables
lmSaturated <- lm(psa ~ ., cancer1)
summary(lmSaturated)
```

```
#rmse
sqrt(mean(lmSaturated$residuals^2))

#10 fold cv
train_control <- trainControl(method = "cv", number = 10)
cv_model <- train(psa ~ ., data = cancer1, method = "lm", trControl =
train_control)
print(cv_model)

#Model Diagnostics
plot(lmSaturated)
#gvlma(lmSaturated)
shapiro.test(resid(lmSaturated))
bptest(lmSaturated)
vif(lmSaturated)

mean(cancer[-c(95,96,97),]$psa)


#Check Potential Outliers
cooks_d <- cooks.distance(lmSaturated)
cooks_d[cooks_d > 0.5]

stu_resid <- studres(lmSaturated)
stu_resid[abs(stu_resid) > 3]

#Remove Outliers
cancer2 <- cancer1[-c(96,97),]
dim(cancer2)

#Saturated Model v2
lmSaturated1 <- lm(psa ~ ., cancer2)
summary(lmSaturated1)

#Model Diagnostics
plot(lmSaturated1)
#gvlma(lmSaturated1)
shapiro.test(resid(lmSaturated1))
bptest(lmSaturated1)
vif(lmSaturated1)

#Check Potential Outliers
cooks_d1 <- cooks.distance(lmSaturated1)
cooks_d1[cooks_d1 > 0.5]

stu_resid1 <- studres(lmSaturated1)
stu_resid1[abs(stu_resid1) > .5]

cancer3 <- cancer2[-c(95,32),]
dim(cancer3)

#Saturated Model v3
#all variables
```

```
lmSaturated2 <- lm(psa ~ ., cancer3)
summary(lmSaturated2)

#rmse
sqrt(mean(lmSaturated2$residuals^2))

#10 fold cv
cv_model1 <- train(psa ~ ., data = cancer3, method = "lm", trControl =
train_control)
print(cv_model1)

plot(lmSaturated2)
#gvlma(lmSaturated2)
hist(resid(lmSaturated2))
shapiro.test(resid(lmSaturated2))
bptest(lmSaturated2)
vif(lmSaturated2)

cooks_d2 <- cooks.distance(lmSaturated2)
cooks_d2[cooks_d2 > 0.5]

stu_resid2 <- studres(lmSaturated2)
stu_resid2[abs(stu_resid2) > 3]

### Naive Models
lmNaive1 <- lm(psa ~ cancerv + weight + age + seminal + capsular + score,
cancer3)
summary(lmNaive1)

#rmse
sqrt(mean(lmSaturated2$residuals^2))

#10 fold cv
cv_model2 <- train(psa ~ cancerv + weight + age + seminal + capsular +
score, data = cancer3, method = "lm", trControl = train_control)
print(cv_model2)

vif(lmNaive1)
plot(lmNaive1)
cooks_d3 <- cooks.distance(lmNaive1)
cooks_d3[cooks_d3 > 0.5]

stu_resid3 <- studres(lmNaive1)
stu_resid3[abs(stu_resid3) > 3]

lmNaive2 <- lm(psa ~ cancerv + weight + seminal + capsular + score,
cancer3)
summary(lmNaive2)

cv_model3 <- train(psa ~ cancerv + weight + seminal + capsular + score,
data = cancer3, method = "lm", trControl = train_control)
print(cv_model3)

plot(lmNaive2)
```

```
cooks_d4 <- cooks.distance(lmNaive2)
cooks_d4[cooks_d4 > 0.4]

stu_resid4 <- studres(lmNaive2)
stu_resid4[abs(stu_resid4) > 3]

anova_compare <- anova(lmNaive2, lmNaive1)
anova_compare


### Ridge

#data prep
#cancer3$seminal <- as.factor(cancer3$seminal)
#cancer3$score <- as.factor(cancer3$score)

y <- cancer3$psa
x <- model.matrix(~ cancerv + weight + age + hyperplasia + seminal +
capsular + score, data = cancer3)[, -1]

set.seed(123)

cv_model_ridge <- cv.glmnet(x, y,
                            alpha = 0,
                            nfolds = 10,
                            standardize = TRUE)

plot(cv_model_ridge)

lambda_min <- cv_model_ridge$lambda.min
lambda_1se <- cv_model_ridge$lambda.1se

best_model <- glmnet(x, y,
                     alpha = 0,
                     lambda = lambda_min,
                     standardize = TRUE)

mse <- mean(cv_model_ridge$cvm[cv_model_ridge$lambda == lambda_min])
rmse <- sqrt(mse)

y_pred <- predict(best_model, newx = x)
r_squared <- 1 - (sum((y - y_pred)^2) / sum((y - mean(y))^2))

n <- length(y)
p <- ncol(x)
adjusted_r_squared <- 1 - (1 - r_squared) * ((n - 1) / (n - p - 1))

summary_info <- list(
  r_squared = r_squared,
  adjusted_r_squared = adjusted_r_squared,
  mse = mse,
  rmse = rmse,
  lambda = lambda_min,
  coefficients = coef(best_model)
```

```
)

print(summary_info)

#10 fold CV on best_model
#10-fold cv
cv_ridge <- train(x = x, y = y,
                  method = "glmnet",
                  trControl = train_control,
                  tuneGrid = expand.grid(alpha = 0, lambda = lambda_min))
print(cv_ridge)

### Lasso
#data prep
y <- cancer3$psa
x <- model.matrix(~ cancerv + weight + age + hyperplasia + seminal +
capsular + score, data = cancer3)[, -1]

set.seed(123)
cv_model_lasso <- cv.glmnet(x, y,
                            alpha = 1,
                            nfolds = 10,
                            standardize = TRUE)

plot(cv_model_lasso)

lambda_min_lasso <- cv_model_lasso$lambda.min
#lambda_min_lasso
lambda_1se_lasso <- cv_model_lasso$lambda.1se
#lambda_1se_lasso

best_model_lasso <- glmnet(x, y,
                           alpha = 1,
                           lambda = lambda_min_lasso,
                           standardize = TRUE)
#coef(best_model_lasso)

#rmse
mse_lasso <- mean(cv_model_lasso$cvm[cv_model_lasso$lambda ==
lambda_min_lasso])
rmse_lasso <- sqrt(mse_lasso)
rmse_lasso

#calc r^2
y_pred_lasso <- predict(best_model_lasso, newx = x)

r_squared_lasso <- 1 - (sum((y - y_pred_lasso)^2) / sum((y - mean(y))^2))

#adj r^2
adjusted_r_squared_lasso <- 1 - (1 - r_squared_lasso) * ((n - 1) / (n - p
- 1))


summary_info_lasso <- list(
```

```
  r_squared = r_squared_lasso,
  adjusted_r_squared = adjusted_r_squared_lasso,
  mse = mse_lasso,
  rmse = rmse_lasso,
  lambda = lambda_min_lasso,
  coefficients = coef(best_model_lasso)
)

print(summary_info_lasso)
#10-fold cv
cv_lasso <- train(x = x, y = y,
                  method = "glmnet",
                  trControl = train_control,
                  tuneGrid = expand.grid(alpha = 1, lambda =
lambda_min_lasso))
print(cv_lasso)

### Lasso without Hyperplasia
#data prep
y1 <- cancer3$psa
x1 <- model.matrix(~ cancerv + weight + age + seminal + capsular + score,
data = cancer3)[, -1]

set.seed(123)
cv_model_lasso1 <- cv.glmnet(x1, y1,
                             alpha = 1,
                             nfolds = 10,
                             standardize = TRUE)

plot(cv_model_lasso1)

lambda_min_lasso1 <- cv_model_lasso1$lambda.min
#lambda_min_lasso 1
lambda_1se_lasso1 <- cv_model_lasso1$lambda.1se
#lambda_1se_lasso1

best_model_lasso1 <- glmnet(x1, y1,
                            alpha = 1,
                            lambda = lambda_min_lasso1,
                            standardize = TRUE)
#coef(best_model_lasso1)

#rmse1
mse_lasso1 <- mean(cv_model_lasso1$cvm[cv_model_lasso1$lambda ==
lambda_min_lasso1])
rmse_lasso1 <- sqrt(mse_lasso1)
rmse_lasso1

#calc r^2
y_pred_lasso1 <- predict(best_model_lasso1, newx = x1)

r_squared_lasso1 <- 1 - (sum((y1 - y_pred_lasso1)^2) / sum((y1 -
mean(y1))^2))
```

```
#adj r^2
n1 <- length(y1)
p1 <- ncol(x1)
adjusted_r_squared_lasso1 <- 1 - (1 - r_squared_lasso1) * ((n1 - 1) / (n1
- p1 - 1))


summary_info_lasso1 <- list(
  r_squared = r_squared_lasso1,
  adjusted_r_squared = adjusted_r_squared_lasso1,
  mse = mse_lasso1,
  rmse = rmse_lasso1,
  lambda = lambda_min_lasso1,
  coefficients = coef(best_model_lasso1)
)

print(summary_info_lasso1)

#10-fold cv
cv_lasso1 <- train(x = x1, y = y1,
                   method = "glmnet",
                   trControl = train_control,
                   tuneGrid = expand.grid(alpha = 1, lambda =
lambda_min_lasso1))
print(cv_lasso1)
```