

ST503 HW1

Halid Kopanski

5/23/2022

Question 1

a)

The data consists of 9 predictors, lcavol, lweight, age, lbph, svi, lcp, gleason, pgg45, lpsa, gleason and age are composed of integers, and svi is binary data.

```
##      lcavol lweight age      lbph svi      lcp gleason pgg45      lpsa
## 1 -0.5798185 2.7695 50 -1.386294 0 -1.38629      6      0 -0.43078
## 2 -0.9942523 3.3196 58 -1.386294 0 -1.38629      6      0 -0.16252
## 3 -0.5108256 2.6912 74 -1.386294 0 -1.38629      7     20 -0.16252
## 4 -1.2039728 3.2828 58 -1.386294 0 -1.38629      6      0 -0.16252
## 5  0.7514161 3.4324 62 -1.386294 0 -1.38629      6      0  0.37156
## 6 -1.0498221 3.2288 50 -1.386294 0 -1.38629      6      0  0.76547
```

```
## 'data.frame':  97 obs. of  9 variables:
## $ lcavol : num  -0.58 -0.994 -0.511 -1.204 0.751 ...
## $ lweight: num  2.77 3.32 2.69 3.28 3.43 ...
## $ age : int  50 58 74 58 62 50 64 58 47 63 ...
## $ lbph : num  -1.39 -1.39 -1.39 -1.39 -1.39 ...
## $ svi : int  0 0 0 0 0 0 0 0 0 0 ...
## $ lcp : num  -1.39 -1.39 -1.39 -1.39 -1.39 ...
## $ gleason: int  6 6 7 6 6 6 6 6 6 ...
## $ pgg45 : int  0 0 20 0 0 0 0 0 0 ...
## $ lpsa : num  -0.431 -0.163 -0.163 -0.163 0.372 ...
```

```
##      lcavol      lweight      age      lbph
## Min.      :-1.3471 Min.      :2.375 Min.      :41.00 Min.      :-1.3863
## 1st Qu.: 0.5128 1st Qu.:3.376 1st Qu.:60.00 1st Qu.: -1.3863
## Median : 1.4469 Median :3.623 Median :65.00 Median : 0.3001
## Mean : 1.3500 Mean :3.653 Mean :63.87 Mean : 0.1004
## 3rd Qu.: 2.1270 3rd Qu.:3.878 3rd Qu.:68.00 3rd Qu.: 1.5581
## Max. : 3.8210 Max. :6.108 Max. :79.00 Max. : 2.3263
##      svi      lcp      gleason      pgg45
## Min.      :0.0000 Min.      :-1.3863 Min.      :6.000 Min.      : 0.00
## 1st Qu.:0.0000 1st Qu.: -1.3863 1st Qu.:6.000 1st Qu.: 0.00
## Median :0.0000 Median : -0.7985 Median :7.000 Median : 15.00
## Mean :0.2165 Mean : -0.1794 Mean :6.753 Mean : 24.38
## 3rd Qu.:0.0000 3rd Qu.: 1.1786 3rd Qu.:7.000 3rd Qu.: 40.00
## Max. :1.0000 Max. : 2.9042 Max. :9.000 Max. :100.00
##      lpsa
```

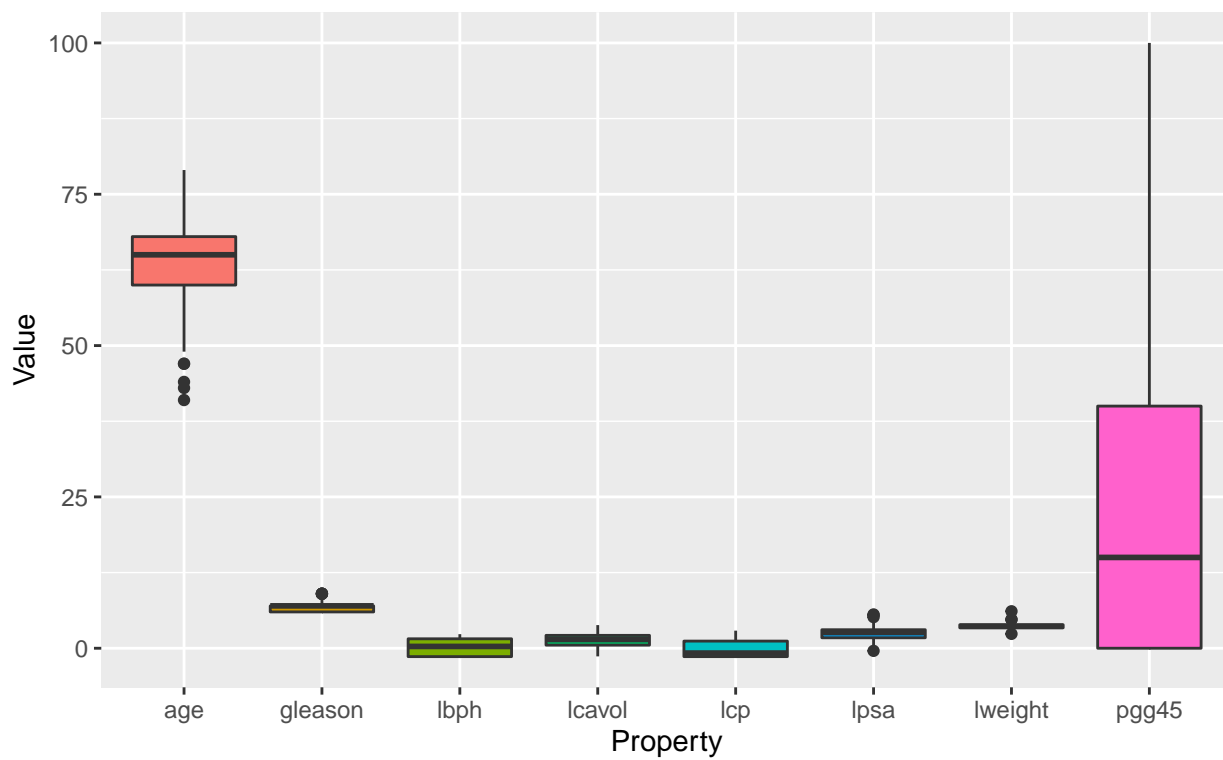
```
## Min.    :-0.4308
## 1st Qu.: 1.7317
## Median : 2.5915
## Mean    : 2.4784
## 3rd Qu.: 3.0564
## Max.    : 5.5829
```

The summary indicates that some of the data is normal or uniform in the cases where median and mean are nearly.

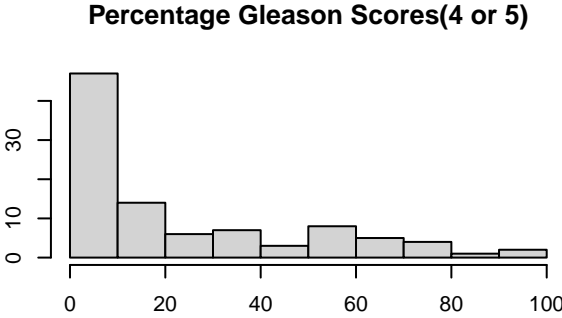
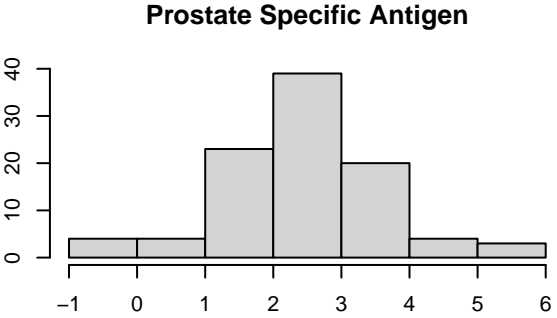
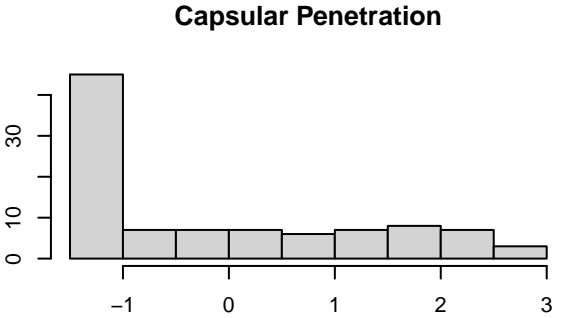
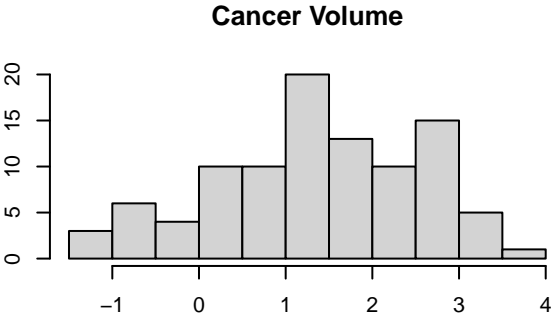
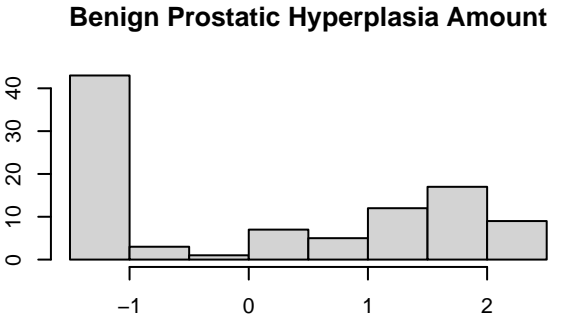
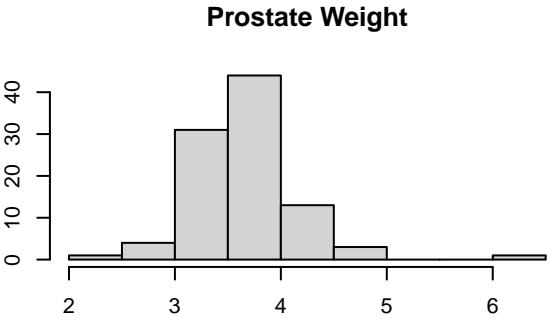
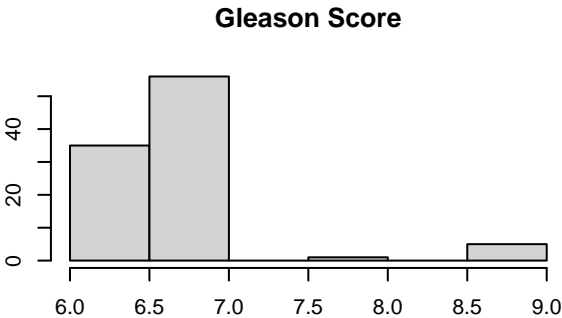
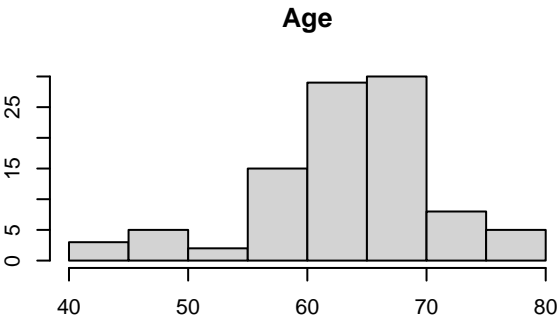
As can be seen in the histograms, the various predictors occupy varying ranges of values. In the cases of age, gleason, lpsa, and lweight a number of outliers can be seen.

Boxplot of Prostate Data

by Halid Kopanski



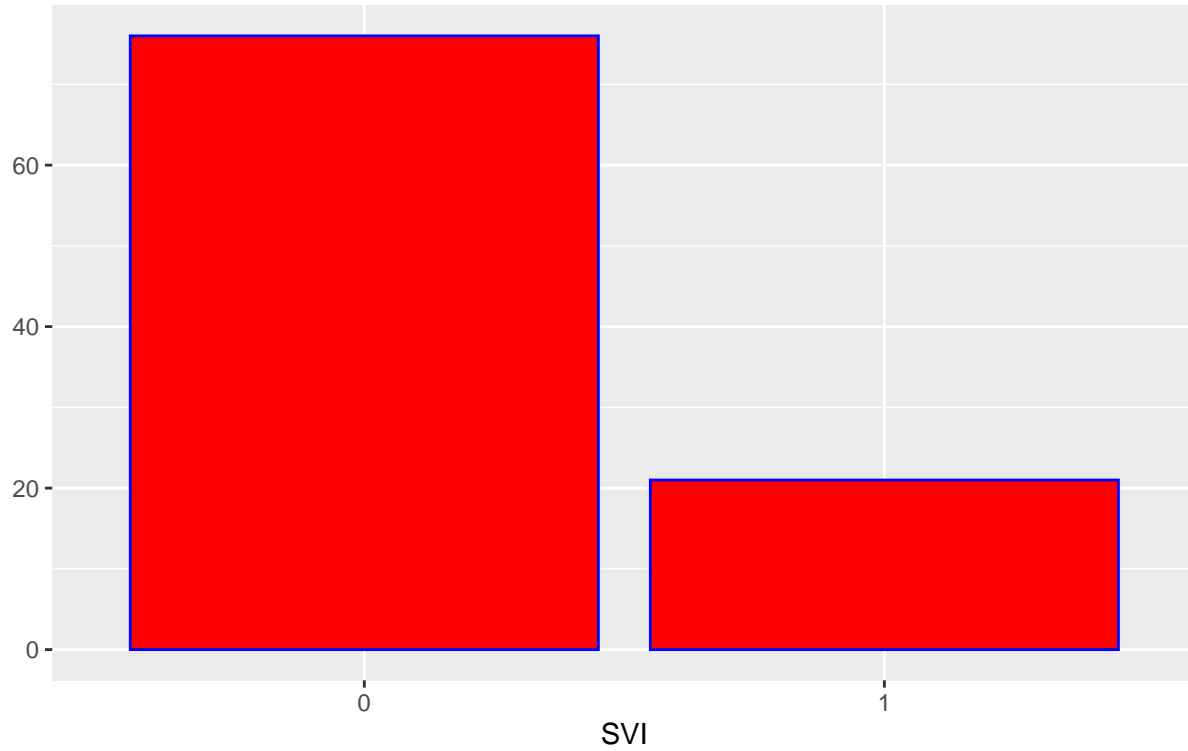
The histograms do indicate that age, prostate weight, cancer weight, and prostate specific antigen are normal like in thier distributions. The remaining predictors do not display any known distributions in first indication.



SVI Value	Value Count
0	76
1	21

Bar Plot of Seminal Vesicle Invasion (SVI)

by Halid Kopanski



b)

In the following table, we can see which predictors had outlier values and how many.

Property	mean	iqr	upper	lower	no_outliers
age	63.8659794	8.000000	48.000000	80.000000	5
gleason	6.7525773	1.000000	4.500000	8.500000	5
lbph	0.1003558	2.944439	-5.802952	5.974804	0
lcavol	1.3500096	1.614217	-1.908502	4.548366	0
lcp	-0.1793637	2.564940	-5.233700	5.026060	0
lpsa	2.4783870	1.324700	-0.255390	5.043410	4
lweight	3.6526887	0.502600	2.622000	4.632400	4
pgg45	24.3814433	40.000000	-60.000000	100.000000	0

Question 4

```
##
## Call:
## lm(formula = gamble ~ ., data = teengamb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -51.082 -11.320  -1.451   9.452  94.252
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.55565   17.19680   1.312   0.1968
## sex         -22.11833    8.21111  -2.694   0.0101 *
## status        0.05223    0.28111   0.186   0.8535
## income        4.96198    1.02539   4.839 1.79e-05 ***
## verbal       -2.95949    2.17215  -1.362   0.1803
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.69 on 42 degrees of freedom
## Multiple R-squared:  0.5267, Adjusted R-squared:  0.4816
## F-statistic: 11.69 on 4 and 42 DF,  p-value: 1.815e-06
```

- a) The R^2 value is 0.5267234. The predictors explain 52.7% of the variance.
- b) Sex has the largest standard error (sum of the residuals) and is calculated to be 8.21
- c) Mean of the residuals is 0 and the median is -1.45.
- d) The correlation between the fitted and the measured values is 0.7257571.
- e) The correlation between the fitted and income level is 0.857142.
- f) Females on average spend 22.12 less on gambling per year than males.

Question 5

a)

```
##
## Call:
## lm(formula = taste ~ ., data = cheddar)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.390  -6.612  -1.009   4.908  25.449
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -28.8768    19.7354  -1.463  0.15540
## Acetic       0.3277     4.4598   0.073  0.94198
## H2S          3.9118     1.2484   3.133  0.00425 **
## Lactic      19.6705     8.6291   2.280  0.03108 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.13 on 26 degrees of freedom
## Multiple R-squared:  0.6518, Adjusted R-squared:  0.6116
## F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

b)

This value is calculated to be 0.6517747 and is equal to the R^2 value.

c)

```
##
## Call:
## lm(formula = taste ~ . + 0, data = cheddar)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.4521  -6.5262  -0.6388   4.6811  28.4744
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## Acetic     -5.454     2.111  -2.583  0.01553 *
## H2S         4.576     1.187   3.854  0.00065 ***
## Lactic     19.127     8.801   2.173  0.03871 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.34 on 27 degrees of freedom
## Multiple R-squared:  0.8877, Adjusted R-squared:  0.8752
## F-statistic: 71.15 on 3 and 27 DF,  p-value: 6.099e-13
```

The R^2 value without the intercept is 0.888. A more reasonable measure would be the adjusted R^2 which was found to be 0.875

d)

$$\begin{array}{r} \hline \hline -28.8767696 \\ 0.3277413 \\ 3.9118411 \\ 19.6705434 \\ \hline \hline \end{array}$$

```

library(faraway)
library(tidyverse)
library(cowplot)
head(prostate)
str(prostate)
summary(prostate)
prostate %>% pivot_longer(-svi, names_to = "Property", values_to = "Values") %>%
ggplot() + geom_boxplot(aes(x = Property, y = Values, fill = Property)) +
labs(x = "Property", y = "Value", title = "Boxplot of Prostate Data",
      subtitle = "by Halid Kopanski") +
theme(legend.position = "none")
#options(repr.plot.width = 6, repr.plot.height = 6, repr.plot.res = 150)
par(mfrow = c(4, 2))
hist(prostate$age, main = "Age", xlab = "", ylab = "")
hist(prostate$gleason, breaks = 8, main = "Gleason Score", xlab = "", ylab = "")
hist(prostate$lweight, main = "Prostate Weight", xlab = "", ylab = "")
hist(prostate$lbph, main = "Benign Prostatic Hyperplasia Amount", xlab = "", ylab = "")
hist(prostate$lcavol, main = "Cancer Volume", xlab = "", ylab = "")
hist(prostate$lcp, main = "Capsular Penetration", xlab = "", ylab = "")
hist(prostate$lpsa, main = "Prostate Specific Antigen", xlab = "", ylab = "")
hist(prostate$pgg45, main = "Percentage Gleason Scores(4 or 5)", xlab = "", ylab = "")
knitr::kable(table(prostate$svi), col.names = c("SVI Value", "Value Count"))

ggplot(prostate) + geom_bar(aes(as.factor(svi)), fill = "red", col = "blue") +
labs(title = "Bar Plot of Seminal Vesicle Invasion (SVI)",
      subtitle = "by Halid Kopanski", x = "SVI", y = "")
prostate %>% pivot_longer(-svi, names_to = "Property", values_to = "Values") %>%
group_by(Property) %>% mutate(outlier = !between(Values,
  as.numeric(quantile(Values)[2]) - 1.5 * IQR(Values),
  as.numeric(quantile(Values)[4]) + 1.5 * IQR(Values))) %>%
summarise(mean = mean(Values),
           iqr = IQR(Values),
           upper = as.numeric(quantile(Values)[2]) - 1.5 * IQR(Values),
           lower = as.numeric(quantile(Values)[4]) + 1.5 * IQR(Values),
           no_outliers = sum(outlier)) %>% knitr::kable()
lin_model <- lm(gamble ~ ., data = teengamb)
summary(lin_model)
r_sqr <- summary(lin_model)$r.squared

mu_res <- round(mean(lin_model$residuals), 2)
med_res <- round(median(lin_model$residuals), 2)

cor_fitted_actual <- cor(lin_model$fitted.values, teengamb$gamble)
cor_fitted_income <- cor(lin_model$fitted.values, teengamb$income)
lin_cheddar <- lm(taste ~ ., cheddar)
summary(lin_cheddar)
r_sqr_emp <- cor(lin_cheddar$fitted.values, cheddar$taste)^2
lin_cheddar2 <- lm(taste ~ . + 0, cheddar)
summary(lin_cheddar2)
qrX <- qr(model.matrix(~ Acetic + H2S + Lactic, cheddar))

Qf <- t(qr.Q(qrX)) %*% as.matrix(cheddar$taste)

```



```
param_reg <- backsolve(qr.R(qrX), Qf)
```