STA 4990 Homework 1

Due 2/7/23

Loading Relevant Packages

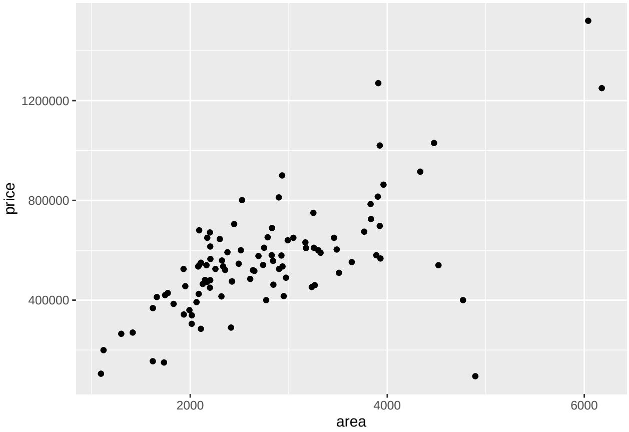
```
# load any relevant packages here, if necessary
library(caret)
library(ggplot2)
library(openintro)
```

Question 4

Part (a)

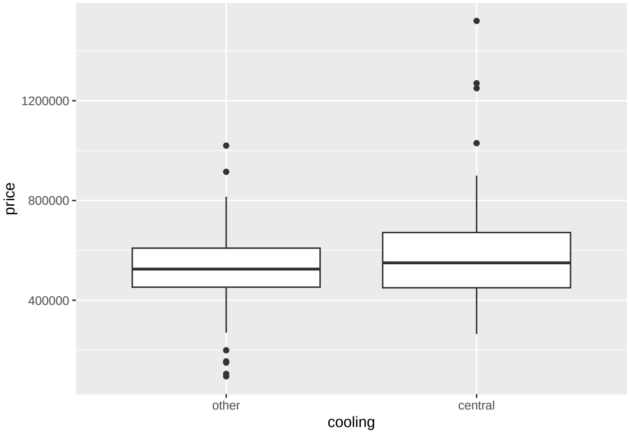
(i)

```
ggplot(duke_forest) + geom_point(aes(x = area , y = price))
```



(ii)

```
ggplot(duke_forest) + geom_boxplot(aes(x = cooling, y = price))
```

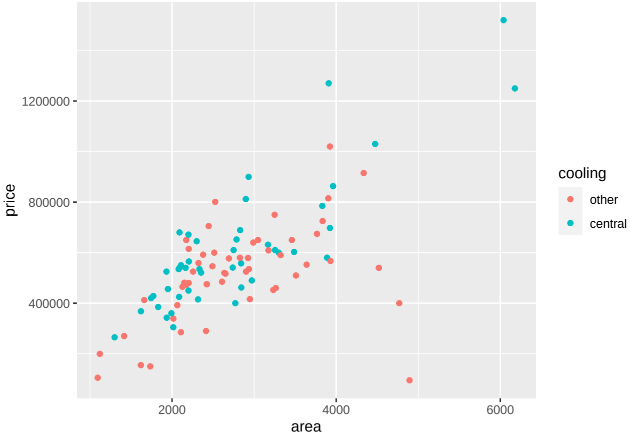


(iii)

What we can infer between the house area and the price is that as the area of the house increases, the price of the house also tends to increase. And what we can infer between the type of cooling system is that the central cooling system tends to be associated with a higher price of the house rather than a different type of colling system.

Part (b)

```
ggplot(duke_forest) + geom_point(aes(x = area, y = price, color = cooling))
```



The color distinction doesn't really affect my interpretation of the relationship between area and price because it still looks like as the area of the house increases the price tends to increase besides some outliers.

Part (c)

(i)

```
m1 <- lm(price ~ area, data = duke_forest)
summary(m1)
```

```
## 
## Call:
## lm(formula = price ~ area, data = duke_forest)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -802163  -70824   -3786   85449  529928 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 116652.33   53302.46   2.188   0.0311 *  
## area           159.48      18.17   8.777 6.29e-14 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 168800 on 96 degrees of freedom
## Multiple R-squared:  0.4452, Adjusted R-squared:  0.4394 
## F-statistic: 77.03 on 1 and 96 DF,  p-value: 6.292e-14
```

(ii)

```
m2 <- glm(price ~ area, data = duke_forest)
summary(m2)
```

```
## 
## Call:
## glm(formula = price ~ area, data = duke_forest)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -802163   -70824    -3786    85449   529928  
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 116652.33   53302.46   2.188   0.0311 *  
## area           159.48      18.17   8.777 6.29e-14 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## (Dispersion parameter for gaussian family taken to be 28492742139)
## 
##     Null deviance: 4.9302e+12  on 97  degrees of freedom
## Residual deviance: 2.7353e+12  on 96  degrees of freedom
## AIC: 2641.2
## 
## Number of Fisher Scoring iterations: 2
```

(iii)

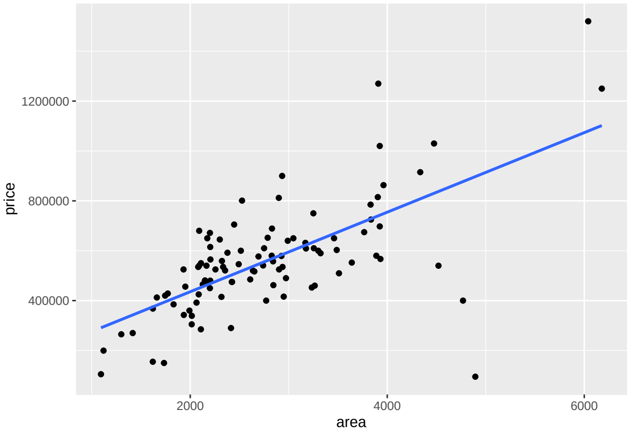
```
m3 <- train(price ~ area, data = duke_forest, method = "lm")
summary(m3)
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -802163  -70824   -3786   85449  529928
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 116652.33   53302.46   2.188   0.0311 *
## area           159.48      18.17   8.777 6.29e-14 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 168800 on 96 degrees of freedom
## Multiple R-squared:  0.4452, Adjusted R-squared:  0.4394
## F-statistic: 77.03 on 1 and 96 DF,  p-value: 6.292e-14
```

The intercept and slope coefficients are the same across all the methods.

Part (d)

```r
ggplot(duke_forest) + geom_point(aes(x = area, y = price)) +
  geom_smooth(method = "lm", aes(x = area, y = price), se = FALSE)
```



I believe that the regression line does NOT appear to fit the data well.

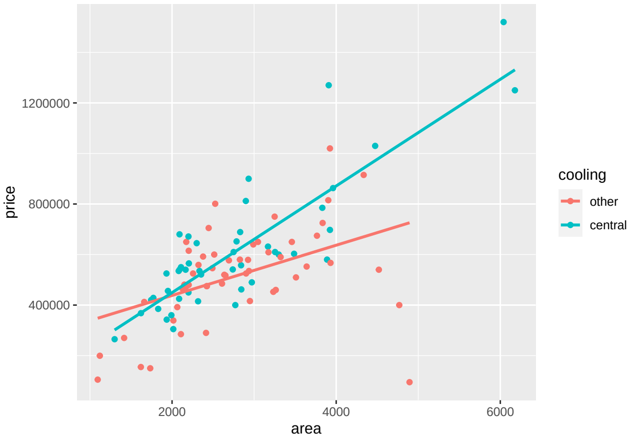
Part (e)

```r
lm(price ~ area * cooling, data = duke_forest)
```

```
##
## Call:
```

```
## lm(formula = price ~ area * cooling, data = duke_forest)
##
## Coefficients:
##         (Intercept)                 area       coolingcentral
##           239255.82                99.41           -211987.70
## area:coolingcentral
##              111.61
```

```
ggplot(duke_forest) + geom_point(aes(x = area, y = price, color = cooling)) +
  geom_smooth(method = "lm", aes(x = area , y = price, color = cooling),
              se = FALSE)
```

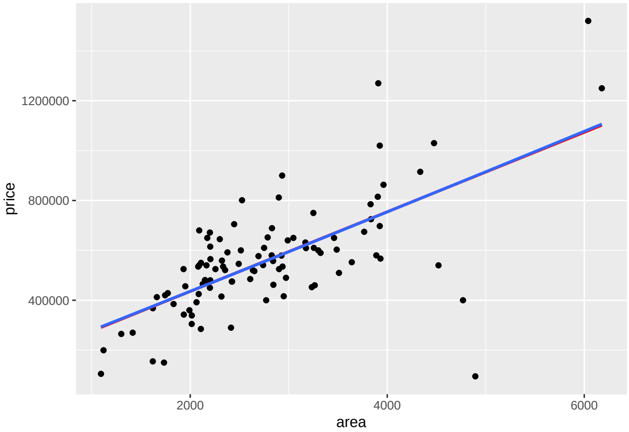


The interaction affects the slope of the line very much, it affected it by the central cooling system having a steeper and much larger slope compared to the slope of a different type of cooling styem. This tells me that on average if the area of the house stayed the same, a house that has a central cooling system will have a higher price compared to a different type of cooling system.

Question 5

Part (a)

```
m_poly2 <- lm(price ~ area + I(area^2), data = duke_forest)
ggplot(duke_forest, aes(x = area, y = price)) +
geom_point() +
geom_smooth(method = "lm", formula = y ~x, se = FALSE, col = "red") +
geom_smooth(method = "lm", formula = y ~ poly(x, 2), se = FALSE)
```



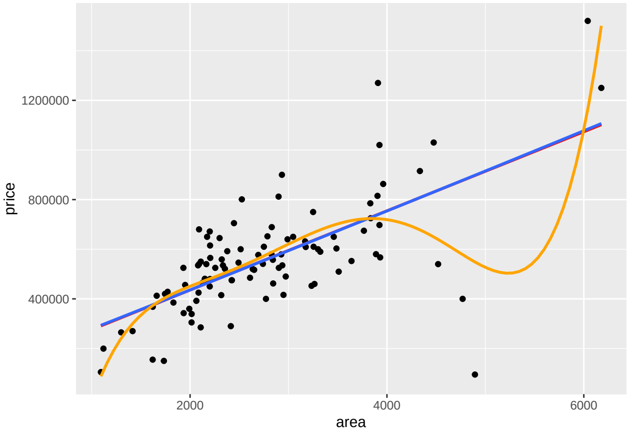
```r
summary(m_poly2)
```

```
##
## Call:
## lm(formula = price ~ area + I(area^2), data = duke_forest)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -803050  -70559   -3196   85378  530414
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.228e+05  1.301e+05   0.944   0.3477
## area        1.553e+02  8.248e+01   1.883   0.0627 .
## I(area^2)   6.308e-04  1.220e-02   0.052   0.9589
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 169700 on 95 degrees of freedom
## Multiple R-squared:  0.4452, Adjusted R-squared:  0.4335
## F-statistic: 38.12 on 2 and 95 DF,  p-value: 7.016e-13
```

I believe that the added flexibility is not necessary because as you can see in the graph the slope and the multiple R-squared did not change much if at all.

Part (b)

```r
m_poly5 <- lm(price ~ poly(area, 5), data = duke_forest)
ggplot(duke_forest, aes(x = area, y = price)) +
geom_point() +
geom_smooth(method = "lm", formula = y ~x, se = FALSE, col = "red") +
geom_smooth(method = "lm", formula = y ~ poly(x, 2), se = FALSE) +
geom_smooth(method = "lm", formula = y ~ poly(x, 5), se = FALSE, col = "orange")
```



I notice that the line of best fit goes through areas that there isnt even points. One reason that this 5th degree polynomial does not make since is that theres no need for it, our points mostly go through a straight line.

Question 6

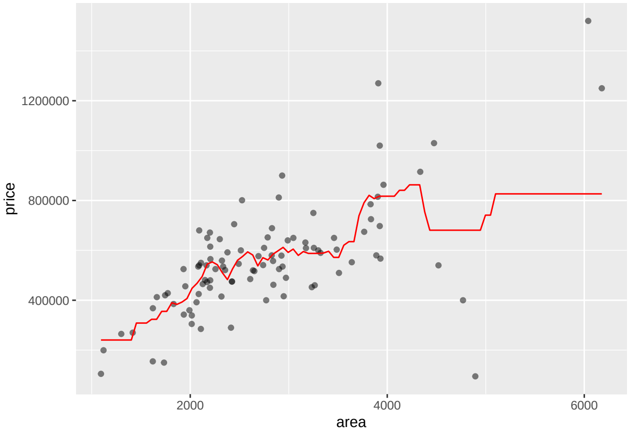
Part (a)

```
library(caret)
set.seed(123) # Ensure reproducibility
knn_fit_8 <- knnreg(price ~ area, data = duke_forest, k = 8)

# Create a sequence of 'area' values for prediction
area_seq <- seq(min(duke_forest$area), max(duke_forest$area), length.out = 100)
pred_data <- data.frame(area = area_seq)

# Predict prices using the kNN model with k = 8
preds_8 <- predict(knn_fit_8, newdata = pred_data)
pred_data$price_8 <- preds_8

# Plotting the kNN prediction for k = 8
ggplot(duke_forest, aes(x = area, y = price)) +
geom_point(alpha = 0.5) +
geom_line(data = pred_data, aes(x = area, y = price_8), color = "red")
```



I feel like it does fit it very well but as the area increases there is a lot of uncertainty about the prediction of the price of the house. But where most of the data is, it does fit it well and you can predict the price of a house in that county.

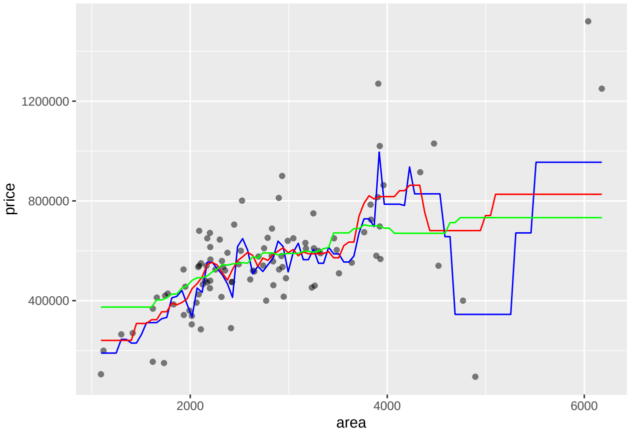
Part (b)

```
# Fit additional kNN models
knn_fit_3 <- knnreg(price ~ area, data = duke_forest, k = 3)
knn_fit_25 <- knnreg(price ~ area, data = duke_forest, k = 25)

# Predict prices using the new kNN models
preds_3 <- predict(knn_fit_3, newdata = pred_data)
preds_25 <- predict(knn_fit_25, newdata = pred_data)

# Add predictions to the data frame for plotting
pred_data$price_3 <- preds_3
pred_data$price_25 <- preds_25

# Update the plot with new kNN predictions
ggplot(duke_forest, aes(x = area, y = price)) +
  geom_point(alpha = 0.5) +
  geom_line(data = pred_data, aes(x = area, y = price_3), color = "blue") +
  geom_line(data = pred_data, aes(x = area, y = price_8), color = "red") +
  geom_line(data = pred_data, aes(x = area, y = price_25), color = "green")
```



The concept of flexibility in the context of kNN is that the smaller value of k leads to a more flexible model, as it focuses more on local patterns, while a larger value of k leads to a less flexible model, as it focuses more on local patterns . In our case a K too little may capture outliers which in return leads to overfitting. And you can clearly see that K = 3 (blue) did capture some outliers towards the end of the graph. While if we choose a K that is too large then that leads to a less flexible model, as it considers a broader range of data points. Bases on the plots I would choose K = 8 because I feel like that is not too small or too large. So the graph may not need to either overfitting or underfitting.