HW3 Complete

```
#Problem 1: The overall goal of this problem is to create the best predictive model for the greenbuildi

library(tidyverse)
```

```
## -- Attaching packages ----------------------- tidyverse 1.3.0 --

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts -------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(FNN)

greenbuildings <- read.csv("~/GitHub/Class Folder/SDS323/data/greenbuildings.csv")
green <- subset(greenbuildings, green_rating==1)
not_green <- subset(greenbuildings, green_rating==0)

#In order to build a predictive model, I first ran KNN regression to fit a linear model to predict the

N = nrow(green)
N_train = floor(0.8*N)
N_test = N - N_train

train_ind = sort(sample.int(N, N_train, replace=FALSE))

D_train = green[train_ind,]
D_train = arrange(D_train, size)
D_test = green[-train_ind,]

y_train = D_train$Rent
X_train = data.frame(size=jitter(D_train$size))
X_test = data.frame(size=jitter(D_test$size))
y_test = D_test$Rent

lm1 = lm(Rent ~ size, data=D_train)
lm2 = lm(Rent ~ poly(size, 2), data=D_train)
knn2 = knn.reg(train = X_train, test = X_test, y = y_train, k=2)

#rmse calculation
rmse = function(y, ypred) {
```

```
  sqrt(mean(data.matrix((y-ypred)^2)))
}

ypred_lm1 = predict(lm1, X_test)
ypred_lm2 = predict(lm2, X_test)
ypred_knn2 = knn2$pred

rmse(y_test, ypred_lm1)
```

```
## [1] 13.51571
```

```
rmse(y_test, ypred_lm2)
```

```
## [1] 13.51777
```

```
rmse(y_test, ypred_knn2)
```

```
## [1] 17.64717
```

```
#attach predictions to data frame
D_test$ypred_lm2 = ypred_lm2
D_test$ypred_knn2 = ypred_knn2

p_test = ggplot(data = D_test) +
  geom_point(mapping = aes(x = size, y = Rent), color='lightgrey') +
  labs(title="Figure 1")
  theme_bw(base_size=18)
```

```
## List of 65
##  $ line                      :List of 6
##   ..$ colour       : chr "black"
##   ..$ size         : num 0.818
##   ..$ linetype     : num 1
##   ..$ lineend      : chr "butt"
##   ..$ arrow        : logi FALSE
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_line" "element"
##  $ rect                      :List of 5
##   ..$ fill         : chr "white"
##   ..$ colour       : chr "black"
##   ..$ size         : num 0.818
##   ..$ linetype     : num 1
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_rect" "element"
##  $ text                      :List of 11
##   ..$ family       : chr ""
##   ..$ face         : chr "plain"
##   ..$ colour       : chr "black"
##   ..$ size         : num 18
##   ..$ hjust        : num 0.5
##   ..$ vjust        : num 0.5
```

```
##   ..$ angle        : num 0
##   ..$ lineheight   : num 0.9
##   ..$ margin       : 'margin' num [1:4] 0pt 0pt 0pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : logi FALSE
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.title.x              :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
##   ..$ vjust        : num 1
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 4.5pt 0pt 0pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.title.x.top          :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
##   ..$ vjust        : num 0
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 0pt 0pt 4.5pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.title.y              :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
##   ..$ vjust        : num 1
##   ..$ angle        : num 90
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 0pt 4.5pt 0pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.title.y.right        :List of 11
```

```
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
##   ..$ vjust        : num 0
##   ..$ angle        : num -90
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 0pt 0pt 0pt 4.5pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.text                 :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : chr "grey30"
##   ..$ size         : 'rel' num 0.8
##   ..$ hjust        : NULL
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.text.x               :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
##   ..$ vjust        : num 1
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 3.6pt 0pt 0pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.text.x.top           :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
##   ..$ vjust        : num 0
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 0pt 0pt 3.6pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
```

```
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.text.y               :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 1
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 0pt 3.6pt 0pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.text.y.right         :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 0
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 0pt 0pt 0pt 3.6pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.ticks                :List of 6
##   ..$ colour       : chr "grey20"
##   ..$ size         : NULL
##   ..$ linetype     : NULL
##   ..$ lineend      : NULL
##   ..$ arrow        : logi FALSE
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_line" "element"
##  $ axis.ticks.length         : 'unit' num 4.5pt
##   ..- attr(*, "valid.unit")= int 8
##   ..- attr(*, "unit")= chr "pt"
##  $ axis.ticks.length.x       : NULL
##  $ axis.ticks.length.x.top   : NULL
##  $ axis.ticks.length.x.bottom: NULL
##  $ axis.ticks.length.y       : NULL
##  $ axis.ticks.length.y.left  : NULL
##  $ axis.ticks.length.y.right : NULL
##  $ axis.line                 : list()
##   ..- attr(*, "class")= chr [1:2] "element_blank" "element"
##  $ axis.line.x               : NULL
##  $ axis.line.y               : NULL
```

```
##  $ legend.background         :List of 5
##   ..$ fill         : NULL
##   ..$ colour       : logi NA
##   ..$ size         : NULL
##   ..$ linetype     : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_rect" "element"
##  $ legend.margin             : 'margin' num [1:4] 9pt 9pt 9pt 9pt
##   ..- attr(*, "valid.unit")= int 8
##   ..- attr(*, "unit")= chr "pt"
##  $ legend.spacing            : 'unit' num 18pt
##   ..- attr(*, "valid.unit")= int 8
##   ..- attr(*, "unit")= chr "pt"
##  $ legend.spacing.x          : NULL
##  $ legend.spacing.y          : NULL
##  $ legend.key                :List of 5
##   ..$ fill         : chr "white"
##   ..$ colour       : logi NA
##   ..$ size         : NULL
##   ..$ linetype     : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_rect" "element"
##  $ legend.key.size           : 'unit' num 1.2lines
##   ..- attr(*, "valid.unit")= int 3
##   ..- attr(*, "unit")= chr "lines"
##  $ legend.key.height         : NULL
##  $ legend.key.width          : NULL
##  $ legend.text               :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : 'rel' num 0.8
##   ..$ hjust        : NULL
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ legend.text.align         : NULL
##  $ legend.title              :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 0
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
```

```
##  $ legend.title.align        : NULL
##  $ legend.position           : chr "right"
##  $ legend.direction          : NULL
##  $ legend.justification      : chr "center"
##  $ legend.box                : NULL
##  $ legend.box.margin         : 'margin' num [1:4] 0cm 0cm 0cm 0cm
##   ..- attr(*, "valid.unit")= int 1
##   ..- attr(*, "unit")= chr "cm"
##  $ legend.box.background     : list()
##   ..- attr(*, "class")= chr [1:2] "element_blank" "element"
##  $ legend.box.spacing        : 'unit' num 18pt
##   ..- attr(*, "valid.unit")= int 8
##   ..- attr(*, "unit")= chr "pt"
##  $ panel.background          :List of 5
##   ..$ fill         : chr "white"
##   ..$ colour       : logi NA
##   ..$ size         : NULL
##   ..$ linetype     : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_rect" "element"
##  $ panel.border              :List of 5
##   ..$ fill         : logi NA
##   ..$ colour       : chr "grey20"
##   ..$ size         : NULL
##   ..$ linetype     : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_rect" "element"
##  $ panel.spacing             : 'unit' num 9pt
##   ..- attr(*, "valid.unit")= int 8
##   ..- attr(*, "unit")= chr "pt"
##  $ panel.spacing.x           : NULL
##  $ panel.spacing.y           : NULL
##  $ panel.grid                :List of 6
##   ..$ colour       : chr "grey92"
##   ..$ size         : NULL
##   ..$ linetype     : NULL
##   ..$ lineend      : NULL
##   ..$ arrow        : logi FALSE
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_line" "element"
##  $ panel.grid.minor          :List of 6
##   ..$ colour       : NULL
##   ..$ size         : 'rel' num 0.5
##   ..$ linetype     : NULL
##   ..$ lineend      : NULL
##   ..$ arrow        : logi FALSE
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_line" "element"
##  $ panel.ontop               : logi FALSE
##  $ plot.background           :List of 5
##   ..$ fill         : NULL
##   ..$ colour       : chr "white"
##   ..$ size         : NULL
##   ..$ linetype     : NULL
```

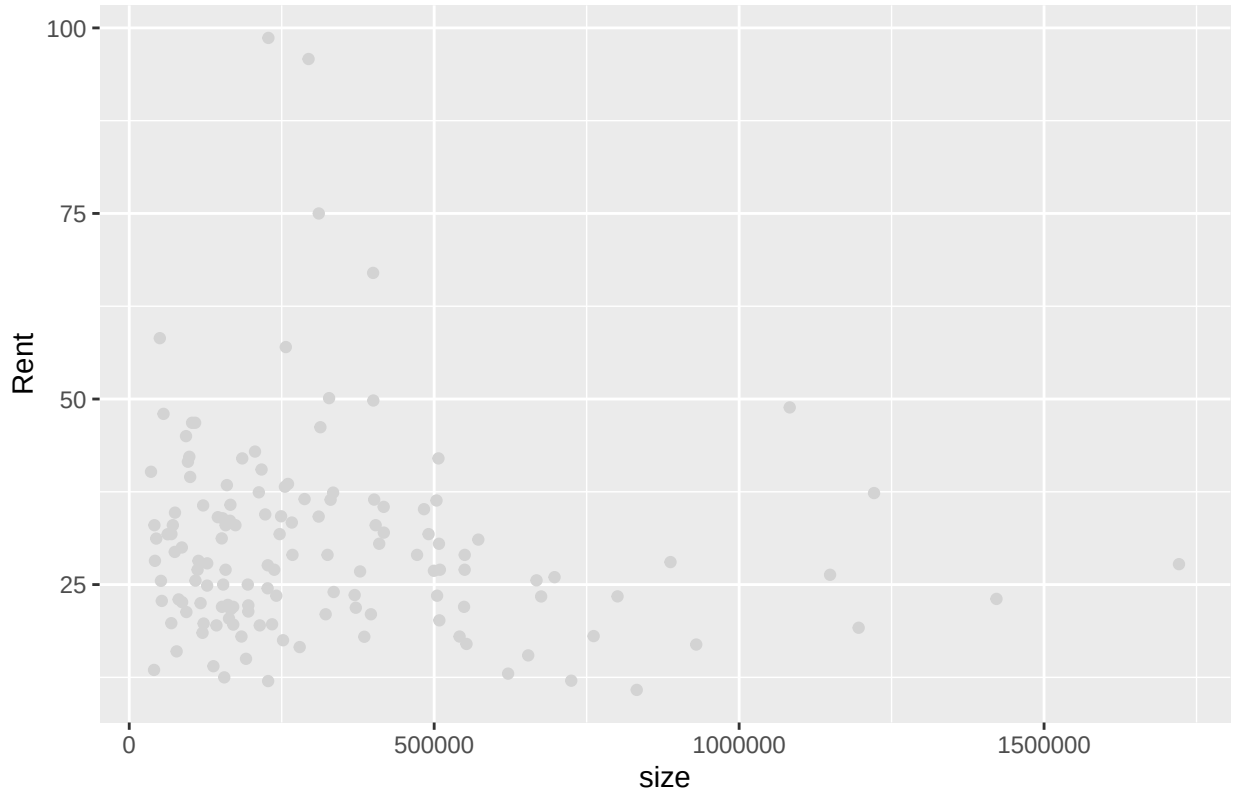
```
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_rect" "element"
##  $ plot.title               :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : 'rel' num 1.2
##   ..$ hjust        : num 0
##   ..$ vjust        : num 1
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 0pt 0pt 9pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ plot.subtitle            :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 0
##   ..$ vjust        : num 1
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 0pt 0pt 9pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ plot.caption             :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : 'rel' num 0.8
##   ..$ hjust        : num 1
##   ..$ vjust        : num 1
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 9pt 0pt 0pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ plot.tag                 :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : 'rel' num 1.2
##   ..$ hjust        : num 0.5
##   ..$ vjust        : num 0.5
```

```
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ plot.tag.position         : chr "topleft"
##  $ plot.margin               : 'margin' num [1:4] 9pt 9pt 9pt 9pt
##   ..- attr(*, "valid.unit")= int 8
##   ..- attr(*, "unit")= chr "pt"
##  $ strip.background          :List of 5
##   ..$ fill         : chr "grey85"
##   ..$ colour       : chr "grey20"
##   ..$ size         : NULL
##   ..$ linetype     : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_rect" "element"
##  $ strip.placement           : chr "inside"
##  $ strip.text                :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : chr "grey10"
##   ..$ size         : 'rel' num 0.8
##   ..$ hjust        : NULL
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 7.2pt 7.2pt 7.2pt 7.2pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ strip.text.x              : NULL
##  $ strip.text.y              :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
##   ..$ vjust        : NULL
##   ..$ angle        : num -90
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ strip.switch.pad.grid     : 'unit' num 4.5pt
##   ..- attr(*, "valid.unit")= int 8
##   ..- attr(*, "unit")= chr "pt"
##  $ strip.switch.pad.wrap     : 'unit' num 4.5pt
##   ..- attr(*, "valid.unit")= int 8
##   ..- attr(*, "unit")= chr "pt"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
```

```
##  - attr(*, "complete")= logi TRUE
##  - attr(*, "validate")= logi TRUE
```

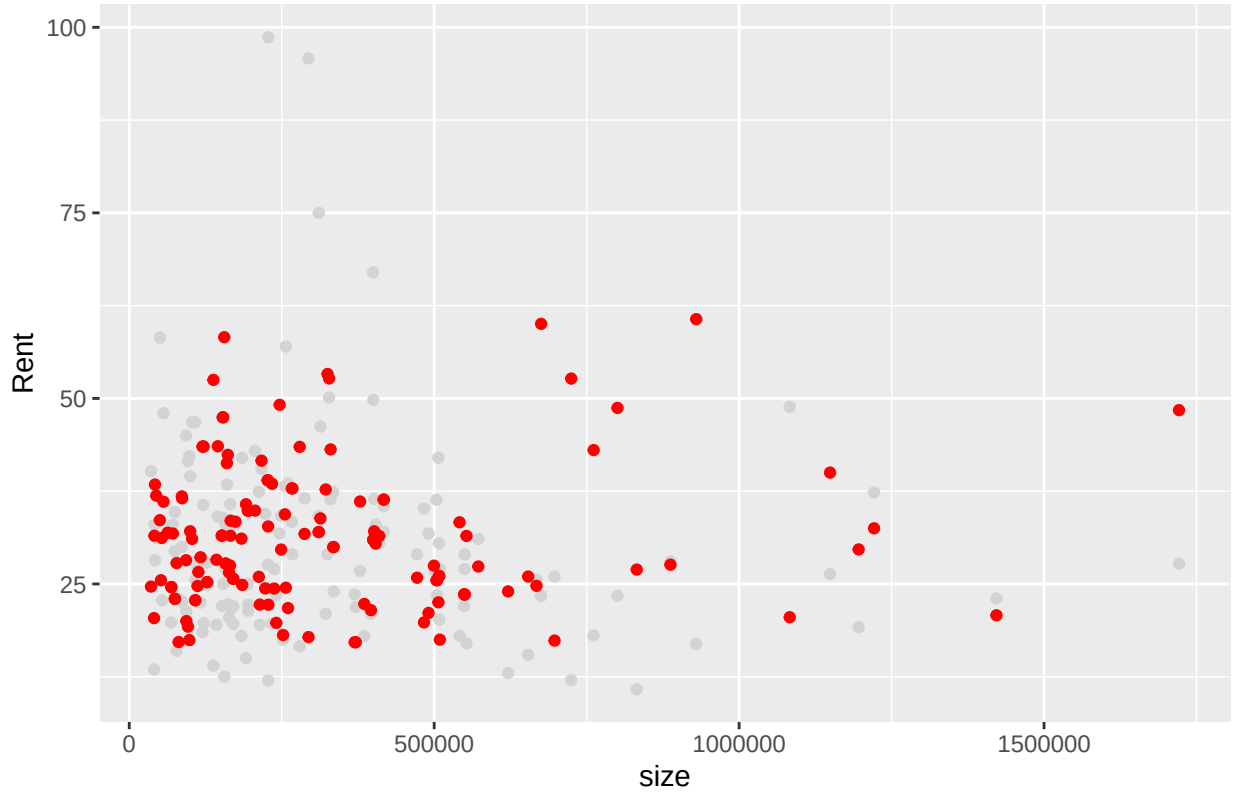
```
p_test
```

Figure 1




```
p_test + geom_point(aes(x = size, y = ypred_knn2), color='red')
```

Figure 1




```
#KNN variances
knn_model = knn.reg(X_train, X_train, y_train, k = 15)

D_train$ypred = knn_model$pred
p_train = ggplot(data = D_train) +
  geom_point(mapping = aes(x = size, y = Rent), color='lightgrey') +
  labs(title="Figure 2")
  theme_bw(base_size=18)
```

```
## List of 65
##  $ line                        :List of 6
##   ..$ colour       : chr "black"
##   ..$ size         : num 0.818
##   ..$ linetype     : num 1
##   ..$ lineend      : chr "butt"
##   ..$ arrow        : logi FALSE
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_line" "element"
##  $ rect                        :List of 5
##   ..$ fill         : chr "white"
##   ..$ colour       : chr "black"
##   ..$ size         : num 0.818
##   ..$ linetype     : num 1
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_rect" "element"
##  $ text                        :List of 11
```

```
##   ..$ family       : chr ""
##   ..$ face         : chr "plain"
##   ..$ colour       : chr "black"
##   ..$ size         : num 18
##   ..$ hjust        : num 0.5
##   ..$ vjust        : num 0.5
##   ..$ angle        : num 0
##   ..$ lineheight   : num 0.9
##   ..$ margin       : 'margin' num [1:4] 0pt 0pt 0pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : logi FALSE
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.title.x              :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
##   ..$ vjust        : num 1
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 4.5pt 0pt 0pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.title.x.top          :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
##   ..$ vjust        : num 0
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 0pt 0pt 4.5pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.title.y              :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
##   ..$ vjust        : num 1
##   ..$ angle        : num 90
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 0pt 4.5pt 0pt 0pt
```

```
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.title.y.right        :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
##   ..$ vjust        : num 0
##   ..$ angle        : num -90
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 0pt 0pt 0pt 4.5pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.text                 :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : chr "grey30"
##   ..$ size         : 'rel' num 0.8
##   ..$ hjust        : NULL
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.text.x               :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
##   ..$ vjust        : num 1
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 3.6pt 0pt 0pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.text.x.top           :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
```

```
##   ..$ vjust        : num 0
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 0pt 0pt 3.6pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.text.y               :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 1
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 0pt 3.6pt 0pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.text.y.right         :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 0
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 0pt 0pt 0pt 3.6pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.ticks                :List of 6
##   ..$ colour       : chr "grey20"
##   ..$ size         : NULL
##   ..$ linetype     : NULL
##   ..$ lineend      : NULL
##   ..$ arrow        : logi FALSE
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_line" "element"
##  $ axis.ticks.length         : 'unit' num 4.5pt
##   ..- attr(*, "valid.unit")= int 8
##   ..- attr(*, "unit")= chr "pt"
##  $ axis.ticks.length.x       : NULL
##  $ axis.ticks.length.x.top   : NULL
##  $ axis.ticks.length.x.bottom: NULL
##  $ axis.ticks.length.y       : NULL
```

```
##  $ axis.ticks.length.y.left  : NULL
##  $ axis.ticks.length.y.right : NULL
##  $ axis.line                 : list()
##   ..- attr(*, "class")= chr [1:2] "element_blank" "element"
##  $ axis.line.x               : NULL
##  $ axis.line.y               : NULL
##  $ legend.background         :List of 5
##   ..$ fill         : NULL
##   ..$ colour       : logi NA
##   ..$ size         : NULL
##   ..$ linetype     : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_rect" "element"
##  $ legend.margin             : 'margin' num [1:4] 9pt 9pt 9pt 9pt
##   ..- attr(*, "valid.unit")= int 8
##   ..- attr(*, "unit")= chr "pt"
##  $ legend.spacing            : 'unit' num 18pt
##   ..- attr(*, "valid.unit")= int 8
##   ..- attr(*, "unit")= chr "pt"
##  $ legend.spacing.x          : NULL
##  $ legend.spacing.y          : NULL
##  $ legend.key                :List of 5
##   ..$ fill         : chr "white"
##   ..$ colour       : logi NA
##   ..$ size         : NULL
##   ..$ linetype     : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_rect" "element"
##  $ legend.key.size           : 'unit' num 1.2lines
##   ..- attr(*, "valid.unit")= int 3
##   ..- attr(*, "unit")= chr "lines"
##  $ legend.key.height         : NULL
##  $ legend.key.width          : NULL
##  $ legend.text               :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : 'rel' num 0.8
##   ..$ hjust        : NULL
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ legend.text.align         : NULL
##  $ legend.title              :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 0
##   ..$ vjust        : NULL
```

```
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ legend.title.align        : NULL
##  $ legend.position           : chr "right"
##  $ legend.direction          : NULL
##  $ legend.justification      : chr "center"
##  $ legend.box                : NULL
##  $ legend.box.margin         : 'margin' num [1:4] 0cm 0cm 0cm 0cm
##   ..- attr(*, "valid.unit")= int 1
##   ..- attr(*, "unit")= chr "cm"
##  $ legend.box.background     : list()
##   ..- attr(*, "class")= chr [1:2] "element_blank" "element"
##  $ legend.box.spacing        : 'unit' num 18pt
##   ..- attr(*, "valid.unit")= int 8
##   ..- attr(*, "unit")= chr "pt"
##  $ panel.background          :List of 5
##   ..$ fill         : chr "white"
##   ..$ colour       : logi NA
##   ..$ size         : NULL
##   ..$ linetype     : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_rect" "element"
##  $ panel.border              :List of 5
##   ..$ fill         : logi NA
##   ..$ colour       : chr "grey20"
##   ..$ size         : NULL
##   ..$ linetype     : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_rect" "element"
##  $ panel.spacing             : 'unit' num 9pt
##   ..- attr(*, "valid.unit")= int 8
##   ..- attr(*, "unit")= chr "pt"
##  $ panel.spacing.x           : NULL
##  $ panel.spacing.y           : NULL
##  $ panel.grid                :List of 6
##   ..$ colour       : chr "grey92"
##   ..$ size         : NULL
##   ..$ linetype     : NULL
##   ..$ lineend      : NULL
##   ..$ arrow        : logi FALSE
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_line" "element"
##  $ panel.grid.minor          :List of 6
##   ..$ colour       : NULL
##   ..$ size         : 'rel' num 0.5
##   ..$ linetype     : NULL
##   ..$ lineend      : NULL
##   ..$ arrow        : logi FALSE
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_line" "element"
```

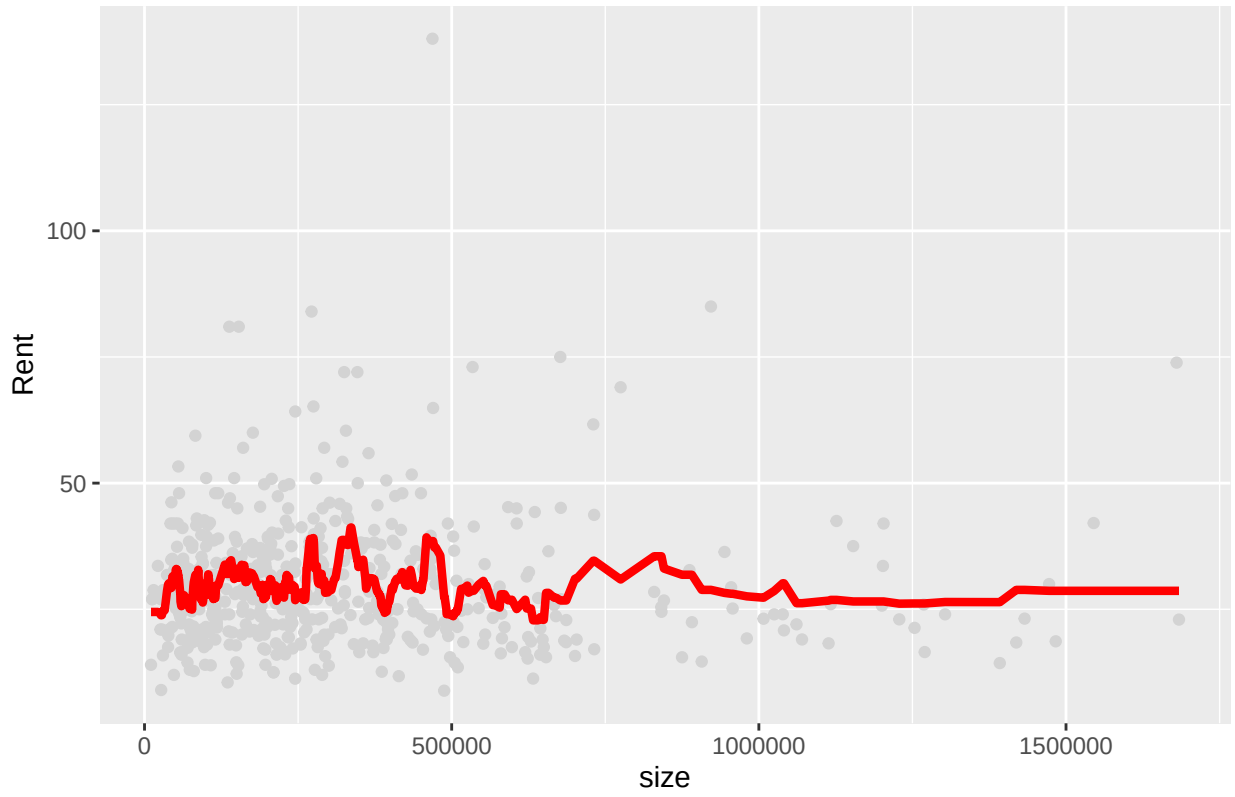
```
##  $ panel.ontop               : logi FALSE
##  $ plot.background           :List of 5
##   ..$ fill         : NULL
##   ..$ colour       : chr "white"
##   ..$ size         : NULL
##   ..$ linetype     : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_rect" "element"
##  $ plot.title                :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : 'rel' num 1.2
##   ..$ hjust        : num 0
##   ..$ vjust        : num 1
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 0pt 0pt 9pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ plot.subtitle             :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 0
##   ..$ vjust        : num 1
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 0pt 0pt 9pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ plot.caption              :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : 'rel' num 0.8
##   ..$ hjust        : num 1
##   ..$ vjust        : num 1
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 9pt 0pt 0pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ plot.tag                  :List of 11
```

```
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : 'rel' num 1.2
##   ..$ hjust        : num 0.5
##   ..$ vjust        : num 0.5
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ plot.tag.position         : chr "topleft"
##  $ plot.margin               : 'margin' num [1:4] 9pt 9pt 9pt 9pt
##   ..- attr(*, "valid.unit")= int 8
##   ..- attr(*, "unit")= chr "pt"
##  $ strip.background          :List of 5
##   ..$ fill         : chr "grey85"
##   ..$ colour       : chr "grey20"
##   ..$ size         : NULL
##   ..$ linetype     : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_rect" "element"
##  $ strip.placement           : chr "inside"
##  $ strip.text                :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : chr "grey10"
##   ..$ size         : 'rel' num 0.8
##   ..$ hjust        : NULL
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 7.2pt 7.2pt 7.2pt 7.2pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ strip.text.x              : NULL
##  $ strip.text.y              :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
##   ..$ vjust        : NULL
##   ..$ angle        : num -90
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ strip.switch.pad.grid     : 'unit' num 4.5pt
```

```
##   ..- attr(*, "valid.unit")= int 8
##   ..- attr(*, "unit")= chr "pt"
##  $ strip.switch.pad.wrap     : 'unit' num 4.5pt
##   ..- attr(*, "valid.unit")= int 8
##   ..- attr(*, "unit")= chr "pt"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi TRUE
##  - attr(*, "validate")= logi TRUE
```

```
p_train + geom_path(mapping = aes(x=size, y=ypred), color='red', size=1.5)
```

Figure 2




```
#These figures show the drawback to using KNN, because size is only one variable affecting the rent. Wh
```

```
#In order to find the best predictive models, forward selection was used to determine the most importan

lm_new <- lm(Rent ~ cluster + size + empl_gr +
               stories + net + amenities+
               hd_total07 + total_dd_07 +
               Precipitation + Gas_Costs + Electricity_Costs +
               cluster_rent, data = greenbuildings)
summary(lm_new)
```

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

```
## lm(formula = Rent ~ cluster + size + empl_gr + stories + net +
##     amenities + hd_total07 + total_dd_07 + Precipitation + Gas_Costs +
##     Electricity_Costs + cluster_rent, data = greenbuildings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.894  -3.712  -0.393   2.642 172.823
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -7.227e+00  9.149e-01  -7.899 3.20e-15 ***
## cluster            8.824e-04  2.849e-04   3.097 0.001964 **
## size               7.887e-06  6.522e-07  12.094  < 2e-16 ***
## empl_gr            5.613e-02  1.696e-02   3.310 0.000937 ***
## stories           -1.836e-02  1.586e-02  -1.157 0.247210
## net               -2.122e+00  5.958e-01  -3.562 0.000370 ***
## amenities          1.406e+00  2.392e-01   5.878 4.32e-09 ***
## hd_total07         5.230e-04  1.339e-04   3.908 9.40e-05 ***
## total_dd_07       -1.383e-04  1.452e-04  -0.953 0.340677
## Precipitation      1.961e-02  1.582e-02   1.240 0.215186
## Gas_Costs         -1.944e+02  7.585e+01  -2.563 0.010386 *
## Electricity_Costs  1.563e+02  2.476e+01   6.311 2.92e-10 ***
## cluster_rent       1.029e+00  1.380e-02  74.572  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.484 on 7807 degrees of freedom
##   (74 observations deleted due to missingness)
## Multiple R-squared:  0.6062, Adjusted R-squared:  0.6056
## F-statistic:  1002 on 12 and 7807 DF,  p-value: < 2.2e-16
```

```
#forward selection process:
lm1 <- lm(Rent ~ cluster, data = greenbuildings)
summary(lm1)
```

```
##
## Call:
## lm(formula = Rent ~ cluster, data = greenbuildings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.818  -8.726  -2.917   5.240 221.748
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.594735   0.297445   82.69   <2e-16 ***
## cluster      0.006496   0.000418   15.54   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.85 on 7892 degrees of freedom
## Multiple R-squared:  0.0297, Adjusted R-squared:  0.02957
## F-statistic: 241.5 on 1 and 7892 DF,  p-value: < 2.2e-16
```

```
lm2 <- lm(Rent ~ cluster + CS_PropertyID, data = greenbuildings)
summary(lm2)
```

```
##
## Call:
## lm(formula = Rent ~ cluster + CS_PropertyID, data = greenbuildings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.769  -8.711  -2.818   5.161 220.845
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.607e+01  3.131e-01   83.25   <2e-16 ***
## cluster        6.334e-03  4.133e-04   15.32   <2e-16 ***
## CS_PropertyID -3.037e-06  2.224e-07  -13.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.68 on 7891 degrees of freedom
## Multiple R-squared:  0.05211,    Adjusted R-squared:  0.05187
## F-statistic: 216.9 on 2 and 7891 DF,  p-value: < 2.2e-16
```

```
lm3 <- lm(Rent ~ cluster + CS_PropertyID + size, data = greenbuildings)
summary(lm3)
```

```
##
## Call:
## lm(formula = Rent ~ cluster + CS_PropertyID + size, data = greenbuildings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.992  -8.884  -2.685   5.386 208.558
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.373e+01  3.541e-01   67.03   <2e-16 ***
## cluster        7.065e-03  4.122e-04   17.14   <2e-16 ***
## CS_PropertyID -2.747e-06  2.209e-07  -12.44   <2e-16 ***
## size           7.545e-06  5.561e-07   13.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.51 on 7890 degrees of freedom
## Multiple R-squared:  0.07372,    Adjusted R-squared:  0.07337
## F-statistic: 209.3 on 3 and 7890 DF,  p-value: < 2.2e-16
```

```
lm4 <- lm(Rent ~ cluster + CS_PropertyID + size + empl_gr, data = greenbuildings)
summary(lm4)
```

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

```
## lm(formula = Rent ~ cluster + CS_PropertyID + size + empl_gr,
##     data = greenbuildings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.810  -8.915  -2.683   5.410 208.605
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.390e+01  3.663e-01   65.23   <2e-16 ***
## cluster        6.989e-03  4.149e-04   16.84   <2e-16 ***
## CS_PropertyID -2.734e-06  2.222e-07  -12.31   <2e-16 ***
## size           7.479e-06  5.595e-07   13.37   <2e-16 ***
## empl_gr       -3.838e-02  2.020e-02   -1.90   0.0575 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.55 on 7815 degrees of freedom
##   (74 observations deleted due to missingness)
## Multiple R-squared:  0.0729, Adjusted R-squared:  0.07243
## F-statistic: 153.6 on 4 and 7815 DF,  p-value: < 2.2e-16
```

```
lm5 <- lm(Rent ~ cluster + CS_PropertyID + size + empl_gr + leasing_rate, data = greenbuildings)
summary(lm5)
```

```
##
## Call:
## lm(formula = Rent ~ cluster + CS_PropertyID + size + empl_gr +
##     leasing_rate, data = greenbuildings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.120  -8.858  -2.676   5.369 209.163
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.622e+01  7.244e-01  22.390   <2e-16 ***
## cluster        6.795e-03  4.113e-04  16.519   <2e-16 ***
## CS_PropertyID -2.425e-06  2.215e-07 -10.948   <2e-16 ***
## size           6.198e-06  5.641e-07  10.988   <2e-16 ***
## empl_gr       -2.914e-02  2.003e-02  -1.455    0.146
## leasing_rate   9.589e-02  7.830e-03  12.247   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.41 on 7814 degrees of freedom
##   (74 observations deleted due to missingness)
## Multiple R-squared:  0.09036,    Adjusted R-squared:  0.08978
## F-statistic: 155.2 on 5 and 7814 DF,  p-value: < 2.2e-16
```

```
lm6 <- lm(Rent ~ cluster + CS_PropertyID + size + empl_gr + leasing_rate
          + stories, data = greenbuildings)
summary(lm6)
```

```
##
## Call:
## lm(formula = Rent ~ cluster + CS_PropertyID + size + empl_gr +
##     leasing_rate + stories, data = greenbuildings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.267  -8.845  -2.671   5.363 209.007
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.625e+01  7.314e-01  22.212  < 2e-16 ***
## cluster        6.789e-03  4.120e-04  16.478  < 2e-16 ***
## CS_PropertyID -2.427e-06  2.216e-07 -10.951  < 2e-16 ***
## size           6.414e-06  9.741e-07   6.584 4.86e-11 ***
## empl_gr       -2.945e-02  2.006e-02  -1.468    0.142
## leasing_rate   9.607e-02  7.858e-03  12.226  < 2e-16 ***
## stories       -6.434e-03  2.369e-02  -0.272    0.786
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.41 on 7813 degrees of freedom
##   (74 observations deleted due to missingness)
## Multiple R-squared:  0.09037,    Adjusted R-squared:  0.08967
## F-statistic: 129.4 on 6 and 7813 DF,  p-value: < 2.2e-16
```

```
lm7 <- lm(Rent ~ cluster + CS_PropertyID + size + empl_gr + leasing_rate
          + stories + age, data = greenbuildings)
summary(lm7)
```

```
##
## Call:
## lm(formula = Rent ~ cluster + CS_PropertyID + size + empl_gr +
##     leasing_rate + stories + age, data = greenbuildings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.555  -8.798  -2.649   5.303 209.739
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.803e+01  8.129e-01  22.179  < 2e-16 ***
## cluster        6.630e-03  4.126e-04  16.069  < 2e-16 ***
## CS_PropertyID -2.449e-06  2.213e-07 -11.067  < 2e-16 ***
## size           5.683e-06  9.836e-07   5.778 7.87e-09 ***
## empl_gr       -3.936e-02  2.013e-02  -1.955   0.0506 .
## leasing_rate   9.233e-02  7.881e-03  11.715  < 2e-16 ***
## stories       -1.422e-03  2.368e-02  -0.060   0.9521
## age           -2.624e-02  5.253e-03  -4.996 5.98e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.39 on 7812 degrees of freedom
##   (74 observations deleted due to missingness)
```

```
## Multiple R-squared:  0.09327,    Adjusted R-squared:  0.09245
## F-statistic: 114.8 on 7 and 7812 DF,  p-value: < 2.2e-16
```

```
lm8 <- lm(Rent ~ cluster + CS_PropertyID + size + empl_gr + leasing_rate
          + stories + age + renovated, data = greenbuildings)
summary(lm8)
```

```
##
## Call:
## lm(formula = Rent ~ cluster + CS_PropertyID + size + empl_gr +
##     leasing_rate + stories + age + renovated, data = greenbuildings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.287  -8.692  -2.403   5.286 211.959
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.811e+01  8.073e-01  22.436  < 2e-16 ***
## cluster        6.407e-03  4.103e-04  15.617  < 2e-16 ***
## CS_PropertyID -2.739e-06  2.215e-07 -12.367  < 2e-16 ***
## size           5.788e-06  9.767e-07   5.926 3.23e-09 ***
## empl_gr       -4.329e-02  1.999e-02  -2.165   0.0304 *
## leasing_rate   9.436e-02  7.828e-03  12.054  < 2e-16 ***
## stories        1.011e-02  2.354e-02   0.430   0.6675
## age            2.924e-03  5.901e-03   0.496   0.6202
## renovated     -4.038e+00  3.821e-01 -10.567  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.29 on 7811 degrees of freedom
##   (74 observations deleted due to missingness)
## Multiple R-squared:  0.106,  Adjusted R-squared:  0.1051
## F-statistic: 115.8 on 8 and 7811 DF,  p-value: < 2.2e-16
```

```
lm9 <- lm(Rent ~ cluster + CS_PropertyID + size + empl_gr + leasing_rate
          + stories + age + renovated + class_a, data = greenbuildings)
summary(lm9)
```

```
##
## Call:
## lm(formula = Rent ~ cluster + CS_PropertyID + size + empl_gr +
##     leasing_rate + stories + age + renovated + class_a, data = greenbuildings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.244  -8.591  -2.387   5.218 212.841
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.605e+01  8.156e-01  19.681  < 2e-16 ***
## cluster        6.378e-03  4.062e-04  15.703  < 2e-16 ***
## CS_PropertyID -2.656e-06  2.193e-07 -12.110  < 2e-16 ***
```

```
## size           4.904e-06  9.695e-07   5.059 4.32e-07 ***
## empl_gr       -4.816e-02  1.980e-02  -2.433   0.0150 *
## leasing_rate   8.571e-02  7.780e-03  11.017  < 2e-16 ***
## stories       -5.266e-02  2.383e-02  -2.210   0.0271 *
## age            3.936e-02  6.515e-03   6.041 1.60e-09 ***
## renovated     -4.137e+00  3.784e-01 -10.934  < 2e-16 ***
## class_a        5.394e+00  4.269e-01  12.633  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.14 on 7810 degrees of freedom
##   (74 observations deleted due to missingness)
## Multiple R-squared:  0.1239, Adjusted R-squared:  0.1229
## F-statistic: 122.8 on 9 and 7810 DF,  p-value: < 2.2e-16
```

```
lm10 <- lm(Rent ~ cluster + CS_PropertyID + size + empl_gr + leasing_rate
          + stories + age + renovated + class_a + class_b, data = greenbuildings)
summary(lm10)
```

```
##
## Call:
## lm(formula = Rent ~ cluster + CS_PropertyID + size + empl_gr +
##     leasing_rate + stories + age + renovated + class_a + class_b,
##     data = greenbuildings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.711  -8.507  -2.285   5.206 213.149
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.473e+01  8.969e-01  16.425  < 2e-16 ***
## cluster        6.357e-03  4.059e-04  15.662  < 2e-16 ***
## CS_PropertyID -2.629e-06  2.193e-07 -11.985  < 2e-16 ***
## size           4.827e-06  9.690e-07   4.981 6.45e-07 ***
## empl_gr       -4.883e-02  1.978e-02  -2.468 0.013599 *
## leasing_rate   8.198e-02  7.846e-03  10.449  < 2e-16 ***
## stories       -5.645e-02  2.383e-02  -2.368 0.017890 *
## age            4.501e-02  6.704e-03   6.714 2.03e-11 ***
## renovated     -4.261e+00  3.797e-01 -11.222  < 2e-16 ***
## class_a        7.036e+00  6.312e-01  11.147  < 2e-16 ***
## class_b        1.799e+00  5.096e-01   3.531 0.000416 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.13 on 7809 degrees of freedom
##   (74 observations deleted due to missingness)
## Multiple R-squared:  0.1253, Adjusted R-squared:  0.1242
## F-statistic: 111.9 on 10 and 7809 DF,  p-value: < 2.2e-16
```

```
lm11 <- lm(Rent ~ cluster + CS_PropertyID + size + empl_gr + leasing_rate
           + stories + age + renovated + class_a + class_b + green_rating, data = greenbuildings)
summary(lm11)
```

```
##
## Call:
## lm(formula = Rent ~ cluster + CS_PropertyID + size + empl_gr +
##     leasing_rate + stories + age + renovated + class_a + class_b +
##     green_rating, data = greenbuildings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.022  -8.510  -2.264   5.200 212.709
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.482e+01  8.967e-01  16.526  < 2e-16 ***
## cluster        6.377e-03  4.057e-04  15.719  < 2e-16 ***
## CS_PropertyID -2.631e-06  2.192e-07 -12.005  < 2e-16 ***
## size           5.021e-06  9.702e-07   5.175 2.33e-07 ***
## empl_gr       -4.939e-02  1.977e-02  -2.498 0.012500 *
## leasing_rate   8.325e-02  7.850e-03  10.604  < 2e-16 ***
## stories       -6.380e-02  2.392e-02  -2.667 0.007672 **
## age            4.324e-02  6.722e-03   6.433 1.33e-10 ***
## renovated     -4.277e+00  3.795e-01 -11.269  < 2e-16 ***
## class_a        7.294e+00  6.356e-01  11.475  < 2e-16 ***
## class_b        1.821e+00  5.093e-01   3.576 0.000351 ***
## green_rating  -1.962e+00  5.933e-01  -3.308 0.000945 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.12 on 7808 degrees of freedom
##   (74 observations deleted due to missingness)
## Multiple R-squared:  0.1266, Adjusted R-squared:  0.1253
## F-statistic: 102.9 on 11 and 7808 DF,  p-value: < 2.2e-16
```

```
lm12 <- lm(Rent ~ cluster + CS_PropertyID + size + empl_gr + leasing_rate
           + stories + age + renovated + class_a + class_b + green_rating + net, data = greenbuildings)
summary(lm12)
```

```
##
## Call:
## lm(formula = Rent ~ cluster + CS_PropertyID + size + empl_gr +
##     leasing_rate + stories + age + renovated + class_a + class_b +
##     green_rating + net, data = greenbuildings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.236  -8.350  -2.236   5.139 211.918
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.495e+01  8.942e-01  16.713  < 2e-16 ***
## cluster        6.382e-03  4.045e-04  15.778  < 2e-16 ***
## CS_PropertyID -2.574e-06  2.187e-07 -11.769  < 2e-16 ***
## size           5.272e-06  9.680e-07   5.447 5.28e-08 ***
## empl_gr       -5.009e-02  1.971e-02  -2.541 0.011063 *
## leasing_rate   8.318e-02  7.827e-03  10.627  < 2e-16 ***
```

```
## stories       -6.127e-02  2.385e-02  -2.569 0.010230 *
## age            4.157e-02  6.706e-03   6.199 5.98e-10 ***
## renovated     -4.252e+00  3.784e-01 -11.236  < 2e-16 ***
## class_a        7.401e+00  6.339e-01  11.675  < 2e-16 ***
## class_b        1.810e+00  5.078e-01   3.565 0.000366 ***
## green_rating  -1.927e+00  5.915e-01  -3.259 0.001125 **
## net           -6.053e+00  8.758e-01  -6.911 5.18e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.08 on 7807 degrees of freedom
##   (74 observations deleted due to missingness)
## Multiple R-squared:  0.1319, Adjusted R-squared:  0.1305
## F-statistic: 98.83 on 12 and 7807 DF,  p-value: < 2.2e-16
```

```
lm13 <- lm(Rent ~ cluster + CS_PropertyID + size + empl_gr + leasing_rate
          + stories + age + renovated + class_a + class_b + green_rating + net +
            amenities, data = greenbuildings)
summary(lm13)
```

```
##
## Call:
## lm(formula = Rent ~ cluster + CS_PropertyID + size + empl_gr +
##     leasing_rate + stories + age + renovated + class_a + class_b +
##     green_rating + net + amenities, data = greenbuildings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.214  -8.374  -2.265   5.118 211.308
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.537e+01  8.961e-01  17.155  < 2e-16 ***
## cluster        6.290e-03  4.041e-04  15.565  < 2e-16 ***
## CS_PropertyID -2.721e-06  2.200e-07 -12.367  < 2e-16 ***
## size           5.745e-06  9.702e-07   5.922 3.32e-09 ***
## empl_gr       -5.097e-02  1.968e-02  -2.590  0.00961 **
## leasing_rate   8.590e-02  7.829e-03  10.972  < 2e-16 ***
## stories       -5.160e-02  2.388e-02  -2.161  0.03072 *
## age            3.894e-02  6.712e-03   5.801 6.85e-09 ***
## renovated     -4.134e+00  3.784e-01 -10.927  < 2e-16 ***
## class_a        8.031e+00  6.435e-01  12.481  < 2e-16 ***
## class_b        2.029e+00  5.085e-01   3.991 6.64e-05 ***
## green_rating  -1.835e+00  5.907e-01  -3.107  0.00190 **
## net           -6.134e+00  8.743e-01  -7.016 2.47e-12 ***
## amenities     -1.995e+00  3.699e-01  -5.393 7.13e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.06 on 7806 degrees of freedom
##   (74 observations deleted due to missingness)
## Multiple R-squared:  0.1351, Adjusted R-squared:  0.1337
## F-statistic:  93.8 on 13 and 7806 DF,  p-value: < 2.2e-16
```

```
lm14 <- lm(Rent ~ cluster + CS_PropertyID + size + empl_gr + leasing_rate
           + stories + age + renovated + class_a + class_b + green_rating + net +
             amenities + cd_total_07, data = greenbuildings)
summary(lm14)
```

```
##
## Call:
## lm(formula = Rent ~ cluster + CS_PropertyID + size + empl_gr +
##     leasing_rate + stories + age + renovated + class_a + class_b +
##     green_rating + net + amenities + cd_total_07, data = greenbuildings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.126  -8.149  -2.236   4.573 211.217
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.922e+01  9.371e-01  20.511  < 2e-16 ***
## cluster        5.662e-03  4.030e-04  14.048  < 2e-16 ***
## CS_PropertyID -2.298e-06  2.203e-07 -10.431  < 2e-16 ***
## size           5.996e-06  9.605e-07   6.243 4.53e-10 ***
## empl_gr        6.999e-02  2.167e-02   3.230  0.00124 **
## leasing_rate   7.993e-02  7.764e-03  10.295  < 2e-16 ***
## stories       -4.760e-02  2.364e-02  -2.014  0.04406 *
## age            2.159e-02  6.782e-03   3.184  0.00146 **
## renovated     -3.955e+00  3.748e-01 -10.553  < 2e-16 ***
## class_a        7.241e+00  6.399e-01  11.314  < 2e-16 ***
## class_b        1.573e+00  5.046e-01   3.118  0.00183 **
## green_rating  -1.648e+00  5.849e-01  -2.818  0.00484 **
## net           -4.939e+00  8.705e-01  -5.673 1.45e-08 ***
## amenities     -1.705e+00  3.668e-01  -4.647 3.42e-06 ***
## cd_total_07   -2.120e-03  1.665e-04 -12.731  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.91 on 7805 degrees of freedom
##   (74 observations deleted due to missingness)
## Multiple R-squared:  0.1527, Adjusted R-squared:  0.1512
## F-statistic: 100.5 on 14 and 7805 DF,  p-value: < 2.2e-16
```

```
lm14 <- lm(Rent ~ cluster + CS_PropertyID + size + empl_gr + leasing_rate
           + stories + age + renovated + class_a + class_b + green_rating + net +
             amenities + cd_total_07 + hd_total07, data = greenbuildings)
summary(lm14)
```

```
##
## Call:
## lm(formula = Rent ~ cluster + CS_PropertyID + size + empl_gr +
##     leasing_rate + stories + age + renovated + class_a + class_b +
##     green_rating + net + amenities + cd_total_07 + hd_total07,
##     data = greenbuildings)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.838  -7.691  -2.480   3.878 211.533
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.443e+01  9.684e-01  25.223  < 2e-16 ***
## cluster        4.291e-03  4.036e-04  10.633  < 2e-16 ***
## CS_PropertyID -2.020e-06  2.168e-07  -9.314  < 2e-16 ***
## size           6.702e-06  9.438e-07   7.101 1.35e-12 ***
## empl_gr        6.729e-02  2.127e-02   3.163 0.001567 **
## leasing_rate   7.959e-02  7.621e-03  10.443  < 2e-16 ***
## stories       -4.826e-03  2.334e-02  -0.207 0.836161
## age            3.894e-02  6.733e-03   5.783 7.61e-09 ***
## renovated     -4.251e+00  3.683e-01 -11.542  < 2e-16 ***
## class_a        6.694e+00  6.290e-01  10.642  < 2e-16 ***
## class_b        1.865e+00  4.956e-01   3.764 0.000168 ***
## green_rating  -2.044e+00  5.746e-01  -3.557 0.000377 ***
## net           -4.777e+00  8.546e-01  -5.589 2.35e-08 ***
## amenities     -1.224e+00  3.612e-01  -3.388 0.000708 ***
## cd_total_07   -2.844e-03  1.688e-04 -16.850  < 2e-16 ***
## hd_total07    -1.521e-03  8.841e-05 -17.199  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.66 on 7804 degrees of freedom
##   (74 observations deleted due to missingness)
## Multiple R-squared:  0.1836, Adjusted R-squared:  0.1821
## F-statistic:   117 on 15 and 7804 DF,  p-value: < 2.2e-16
```

```
lm15 <- lm(Rent ~ cluster + CS_PropertyID + size + empl_gr + leasing_rate
           + stories + age + renovated + class_a + class_b + green_rating + net +
             amenities + cd_total_07 + hd_total07 + total_dd_07, data = greenbuildings)
summary(lm15)
```

```
##
## Call:
## lm(formula = Rent ~ cluster + CS_PropertyID + size + empl_gr +
##     leasing_rate + stories + age + renovated + class_a + class_b +
##     green_rating + net + amenities + cd_total_07 + hd_total07 +
##     total_dd_07, data = greenbuildings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.838  -7.691  -2.480   3.878 211.533
##
## Coefficients: (1 not defined because of singularities)
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.443e+01  9.684e-01  25.223  < 2e-16 ***
## cluster        4.291e-03  4.036e-04  10.633  < 2e-16 ***
## CS_PropertyID -2.020e-06  2.168e-07  -9.314  < 2e-16 ***
## size           6.702e-06  9.438e-07   7.101 1.35e-12 ***
## empl_gr        6.729e-02  2.127e-02   3.163 0.001567 **
## leasing_rate   7.959e-02  7.621e-03  10.443  < 2e-16 ***
```

```
## stories       -4.826e-03  2.334e-02  -0.207 0.836161
## age            3.894e-02  6.733e-03   5.783 7.61e-09 ***
## renovated     -4.251e+00  3.683e-01 -11.542  < 2e-16 ***
## class_a        6.694e+00  6.290e-01  10.642  < 2e-16 ***
## class_b        1.865e+00  4.956e-01   3.764 0.000168 ***
## green_rating  -2.044e+00  5.746e-01  -3.557 0.000377 ***
## net           -4.777e+00  8.546e-01  -5.589 2.35e-08 ***
## amenities     -1.224e+00  3.612e-01  -3.388 0.000708 ***
## cd_total_07   -2.844e-03  1.688e-04 -16.850  < 2e-16 ***
## hd_total07    -1.521e-03  8.841e-05 -17.199  < 2e-16 ***
## total_dd_07           NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.66 on 7804 degrees of freedom
##   (74 observations deleted due to missingness)
## Multiple R-squared:  0.1836, Adjusted R-squared:  0.1821
## F-statistic:   117 on 15 and 7804 DF,  p-value: < 2.2e-16
```

```
lm16 <- lm(Rent ~ cluster + CS_PropertyID + size + empl_gr + leasing_rate
           + stories + age + renovated + class_a + class_b + green_rating + net +
             amenities + cd_total_07 + hd_total07 + total_dd_07 +
             Precipitation, data = greenbuildings)
summary(lm16)
```

```
##
## Call:
## lm(formula = Rent ~ cluster + CS_PropertyID + size + empl_gr +
##     leasing_rate + stories + age + renovated + class_a + class_b +
##     green_rating + net + amenities + cd_total_07 + hd_total07 +
##     total_dd_07 + Precipitation, data = greenbuildings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.996  -7.457  -2.053   4.483 208.036
##
## Coefficients: (1 not defined because of singularities)
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.814e+01  9.890e-01  18.346  < 2e-16 ***
## cluster        4.742e-03  3.933e-04  12.056  < 2e-16 ***
## CS_PropertyID -1.725e-06  2.115e-07  -8.158 3.95e-16 ***
## size           6.788e-06  9.184e-07   7.391 1.61e-13 ***
## empl_gr        2.537e-01  2.253e-02  11.257  < 2e-16 ***
## leasing_rate   7.545e-02  7.419e-03  10.170  < 2e-16 ***
## stories       -2.741e-02  2.273e-02  -1.206  0.22802
## age            2.717e-02  6.576e-03   4.131 3.65e-05 ***
## renovated     -4.279e+00  3.584e-01 -11.938  < 2e-16 ***
## class_a        6.246e+00  6.124e-01  10.198  < 2e-16 ***
## class_b        2.375e+00  4.829e-01   4.919 8.87e-07 ***
## green_rating  -1.489e+00  5.598e-01  -2.660  0.00782 **
## net           -4.437e+00  8.318e-01  -5.334 9.88e-08 ***
## amenities     -9.062e-01  3.518e-01  -2.576  0.01001 *
## cd_total_07   -4.457e-03  1.814e-04 -24.568  < 2e-16 ***
## hd_total07    -2.168e-03  9.142e-05 -23.717  < 2e-16 ***
```

```
## total_dd_07           NA         NA      NA       NA
## Precipitation  3.370e-01  1.609e-02  20.941  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.29 on 7803 degrees of freedom
##   (74 observations deleted due to missingness)
## Multiple R-squared:  0.2271, Adjusted R-squared:  0.2255
## F-statistic: 143.3 on 16 and 7803 DF,  p-value: < 2.2e-16
```

```
lm17 <- lm(Rent ~ cluster + CS_PropertyID + size + empl_gr + leasing_rate
           + stories + age + renovated + class_a + class_b + green_rating + net +
             amenities + cd_total_07 + hd_total07 + total_dd_07 +
             Precipitation + Gas_Costs, data = greenbuildings)
summary(lm17)
```

```
##
## Call:
## lm(formula = Rent ~ cluster + CS_PropertyID + size + empl_gr +
##     leasing_rate + stories + age + renovated + class_a + class_b +
##     green_rating + net + amenities + cd_total_07 + hd_total07 +
##     total_dd_07 + Precipitation + Gas_Costs, data = greenbuildings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.028  -7.498  -2.049   4.483 208.022
##
## Coefficients: (1 not defined because of singularities)
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.727e+01  1.175e+00  14.694  < 2e-16 ***
## cluster        4.786e-03  3.946e-04  12.129  < 2e-16 ***
## CS_PropertyID -1.807e-06  2.196e-07  -8.228  < 2e-16 ***
## size           6.939e-06  9.249e-07   7.503 6.93e-14 ***
## empl_gr        2.535e-01  2.253e-02  11.251  < 2e-16 ***
## leasing_rate   7.453e-02  7.448e-03  10.007  < 2e-16 ***
## stories       -2.917e-02  2.277e-02  -1.281  0.20018
## age            2.699e-02  6.577e-03   4.104 4.10e-05 ***
## renovated     -4.273e+00  3.584e-01 -11.923  < 2e-16 ***
## class_a        6.161e+00  6.155e-01  10.009  < 2e-16 ***
## class_b        2.341e+00  4.835e-01   4.841 1.32e-06 ***
## green_rating  -1.446e+00  5.606e-01  -2.578  0.00994 **
## net           -4.530e+00  8.345e-01  -5.429 5.85e-08 ***
## amenities     -8.593e-01  3.534e-01  -2.432  0.01506 *
## cd_total_07   -4.538e-03  1.907e-04 -23.796  < 2e-16 ***
## hd_total07    -2.140e-03  9.362e-05 -22.861  < 2e-16 ***
## total_dd_07           NA         NA      NA       NA
## Precipitation  3.216e-01  1.959e-02  16.412  < 2e-16 ***
## Gas_Costs      1.293e+02  9.381e+01   1.378  0.16818
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.29 on 7802 degrees of freedom
##   (74 observations deleted due to missingness)
## Multiple R-squared:  0.2273, Adjusted R-squared:  0.2256
```

```
## F-statistic:   135 on 17 and 7802 DF,  p-value: < 2.2e-16
```

```
lm18 <- lm(Rent ~ cluster + CS_PropertyID + size + empl_gr + leasing_rate
           + stories + age + renovated + class_a + class_b + green_rating + net +
             amenities + cd_total_07 + hd_total07 + total_dd_07 +
             Precipitation + Gas_Costs + Electricity_Costs, data = greenbuildings)
summary(lm18)
```

```
##
## Call:
## lm(formula = Rent ~ cluster + CS_PropertyID + size + empl_gr +
##     leasing_rate + stories + age + renovated + class_a + class_b +
##     green_rating + net + amenities + cd_total_07 + hd_total07 +
##     total_dd_07 + Precipitation + Gas_Costs + Electricity_Costs,
##     data = greenbuildings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.485  -6.812  -1.391   4.742 186.403
##
## Coefficients: (1 not defined because of singularities)
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1.309e+01  1.303e+00 -10.049  < 2e-16 ***
## cluster            2.745e-03  3.620e-04   7.584 3.74e-14 ***
## CS_PropertyID     -1.259e-06  1.999e-07  -6.295 3.25e-10 ***
## size               5.673e-06  8.408e-07   6.748 1.61e-11 ***
## empl_gr            4.161e-01  2.086e-02  19.951  < 2e-16 ***
## leasing_rate       6.156e-02  6.774e-03   9.088  < 2e-16 ***
## stories           -7.769e-02  2.072e-02  -3.750 0.000178 ***
## age               -1.167e-02  6.050e-03  -1.929 0.053751 .
## renovated         -2.197e+00  3.296e-01  -6.667 2.79e-11 ***
## class_a            5.487e+00  5.594e-01   9.808  < 2e-16 ***
## class_b            2.116e+00  4.393e-01   4.818 1.48e-06 ***
## green_rating      -9.054e-01  5.095e-01  -1.777 0.075597 .
## net               -5.578e+00  7.585e-01  -7.354 2.11e-13 ***
## amenities          4.689e-01  3.227e-01   1.453 0.146260
## cd_total_07       -1.905e-03  1.850e-04 -10.301  < 2e-16 ***
## hd_total07         9.929e-04  1.148e-04   8.651  < 2e-16 ***
## total_dd_07               NA         NA      NA       NA
## Precipitation      5.439e-01  1.862e-02  29.209  < 2e-16 ***
## Gas_Costs         -1.771e+03  9.720e+01 -18.222  < 2e-16 ***
## Electricity_Costs  1.110e+03  2.730e+01  40.653  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.07 on 7801 degrees of freedom
##   (74 observations deleted due to missingness)
## Multiple R-squared:  0.3624, Adjusted R-squared:  0.3609
## F-statistic: 246.3 on 18 and 7801 DF,  p-value: < 2.2e-16
```

```
lm19 <- lm(Rent ~ cluster + CS_PropertyID + size + empl_gr + leasing_rate
           + stories + age + renovated + class_a + class_b + green_rating + net +
             amenities + cd_total_07 + hd_total07 + total_dd_07 +
```

```
            Precipitation + Gas_Costs + Electricity_Costs +
            cluster_rent, data = greenbuildings)
summary(lm19)
```

```
##
## Call:
## lm(formula = Rent ~ cluster + CS_PropertyID + size + empl_gr +
##     leasing_rate + stories + age + renovated + class_a + class_b +
##     green_rating + net + amenities + cd_total_07 + hd_total07 +
##     total_dd_07 + Precipitation + Gas_Costs + Electricity_Costs +
##     cluster_rent, data = greenbuildings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.765  -3.581  -0.530   2.483 173.892
##
## Coefficients: (1 not defined because of singularities)
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -8.341e+00  1.018e+00  -8.194 2.94e-16 ***
## cluster            7.340e-04  2.836e-04   2.588 0.009677 **
## CS_PropertyID      2.992e-07  1.574e-07   1.901 0.057355 .
## size               6.776e-06  6.557e-07  10.334  < 2e-16 ***
## empl_gr            6.483e-02  1.700e-02   3.814 0.000138 ***
## leasing_rate       9.433e-03  5.332e-03   1.769 0.076880 .
## stories           -3.534e-02  1.616e-02  -2.186 0.028831 *
## age               -1.256e-02  4.717e-03  -2.664 0.007740 **
## renovated         -1.373e-01  2.586e-01  -0.531 0.595310
## class_a            2.872e+00  4.377e-01   6.562 5.66e-11 ***
## class_b            1.191e+00  3.427e-01   3.474 0.000516 ***
## green_rating       6.566e-01  3.978e-01   1.650 0.098895 .
## net               -2.553e+00  5.929e-01  -4.307 1.68e-05 ***
## amenities          6.540e-01  2.516e-01   2.599 0.009354 **
## cd_total_07       -1.328e-04  1.463e-04  -0.908 0.364050
## hd_total07         5.379e-04  8.971e-05   5.996 2.12e-09 ***
## total_dd_07               NA         NA      NA       NA
## Precipitation      4.877e-02  1.611e-02   3.028 0.002474 **
## Gas_Costs         -3.521e+02  7.837e+01  -4.493 7.11e-06 ***
## Electricity_Costs  1.882e+02  2.493e+01   7.548 4.93e-14 ***
## cluster_rent       1.008e+00  1.421e-02  70.956  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.413 on 7800 degrees of freedom
##   (74 observations deleted due to missingness)
## Multiple R-squared:  0.6125, Adjusted R-squared:  0.6115
## F-statistic: 648.9 on 19 and 7800 DF,  p-value: < 2.2e-16
```

```
lm_new <- lm(Rent ~ cluster + size + empl_gr +
               stories + net + amenities+
               hd_total07 + total_dd_07 +
               Precipitation + Gas_Costs + Electricity_Costs +
               cluster_rent, data = greenbuildings)
summary(lm_new)
```

```
##
## Call:
## lm(formula = Rent ~ cluster + size + empl_gr + stories + net +
##     amenities + hd_total07 + total_dd_07 + Precipitation + Gas_Costs +
##     Electricity_Costs + cluster_rent, data = greenbuildings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.894  -3.712  -0.393   2.642 172.823
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -7.227e+00  9.149e-01  -7.899 3.20e-15 ***
## cluster            8.824e-04  2.849e-04   3.097 0.001964 **
## size               7.887e-06  6.522e-07  12.094  < 2e-16 ***
## empl_gr            5.613e-02  1.696e-02   3.310 0.000937 ***
## stories           -1.836e-02  1.586e-02  -1.157 0.247210
## net               -2.122e+00  5.958e-01  -3.562 0.000370 ***
## amenities          1.406e+00  2.392e-01   5.878 4.32e-09 ***
## hd_total07         5.230e-04  1.339e-04   3.908 9.40e-05 ***
## total_dd_07       -1.383e-04  1.452e-04  -0.953 0.340677
## Precipitation      1.961e-02  1.582e-02   1.240 0.215186
## Gas_Costs         -1.944e+02  7.585e+01  -2.563 0.010386 *
## Electricity_Costs  1.563e+02  2.476e+01   6.311 2.92e-10 ***
## cluster_rent       1.029e+00  1.380e-02  74.572  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.484 on 7807 degrees of freedom
##   (74 observations deleted due to missingness)
## Multiple R-squared:  0.6062, Adjusted R-squared:  0.6056
## F-statistic:  1002 on 12 and 7807 DF,  p-value: < 2.2e-16
```

```
#According to this model that only includes the statistically significant variables at a .05 level, the

# Split into training and testing sets
n = nrow(greenbuildings)
n_train = round(0.8*n)
n_test = n - n_train
train_cases = sample.int(n, n_train, replace=FALSE)
test_cases = setdiff(1:n, train_cases)
green_train = greenbuildings[train_cases,]
green_test = greenbuildings[test_cases,]

# Fit to the training data
train_lm19 = lm(Rent ~ cluster + CS_PropertyID + size + empl_gr + leasing_rate
                + stories + age + renovated + class_a + class_b + green_rating + net +
                  amenities + cd_total_07 + hd_total07 + total_dd_07 +
                  Precipitation + Gas_Costs + Electricity_Costs +
                  cluster_rent, data = green_train)
train_lm_new = lm(Rent ~ cluster + size + empl_gr +
                    stories + net + amenities+
                    hd_total07 + total_dd_07 +
                    Precipitation + Gas_Costs + Electricity_Costs +
```

```
                    cluster_rent, data = green_train)
train_lm_new2 = lm(log(Rent) ~ cluster + size + empl_gr +
                             stories + net + amenities+
                             hd_total07 + total_dd_07 +
                             Precipitation + Gas_Costs + Electricity_Costs +
                             cluster_rent, data = green_train)
train_lm_new3 = lm(Rent ~ (cluster + size + empl_gr +
                                 stories + net + amenities+
                                 hd_total07 + total_dd_07 +
                                 Precipitation + Gas_Costs + Electricity_Costs +
                                 cluster_rent)^2, data = green_train)

# Predictions out of sample
ypred1 = predict(train_lm19, green_test)
```

```
## Warning in predict.lm(train_lm19, green_test): prediction from a rank-deficient
## fit may be misleading
```

```
ypred2 = predict(train_lm_new, green_test)
ypred3 = predict(train_lm_new2, green_test)
ypred4 = predict(train_lm_new3, green_test)


summary(ypred1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   8.277  20.431  26.599  28.438  33.978  80.505      14
```

```
summary(ypred2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    7.51   20.46   26.49   28.44   34.50   81.89      14
```

```
summary(ypred3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   2.472   2.988   3.202   3.243   3.466   4.836      14
```

```
summary(ypred4)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   9.732  20.444  25.856  28.378  33.378 103.850      14
```

```
#Based on the above process, the predictive model that will be used to find the price of rent is the lm

#average rent increase with green certification using lm19
lm_green <- lm(Rent ~ cluster + CS_PropertyID + size + empl_gr + leasing_rate
          + stories + age + renovated + class_a + class_b + green_rating + net +
            amenities + cd_total_07 + hd_total07 + total_dd_07 +
            Precipitation + Gas_Costs + Electricity_Costs +
            cluster_rent, data = green)
summary(lm_green)
```

```
##
## Call:
## lm(formula = Rent ~ cluster + CS_PropertyID + size + empl_gr +
##     leasing_rate + stories + age + renovated + class_a + class_b +
##     green_rating + net + amenities + cd_total_07 + hd_total07 +
##     total_dd_07 + Precipitation + Gas_Costs + Electricity_Costs +
##     cluster_rent, data = green)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.707  -3.254  -0.445   2.747  59.486
##
## Coefficients: (2 not defined because of singularities)
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -8.330e+00  3.959e+00  -2.104  0.03575 *
## cluster            1.597e-03  6.911e-04   2.311  0.02115 *
## CS_PropertyID      7.183e-07  5.852e-07   1.228  0.22005
## size               5.554e-06  2.347e-06   2.366  0.01826 *
## empl_gr            7.287e-02  3.564e-02   2.045  0.04128 *
## leasing_rate       3.679e-02  2.093e-02   1.758  0.07920 .
## stories           -2.799e-02  5.067e-02  -0.552  0.58089
## age                1.614e-02  1.999e-02   0.808  0.41964
## renovated         -3.142e-01  6.898e-01  -0.456  0.64886
## class_a            1.260e+00  2.645e+00   0.477  0.63385
## class_b            5.121e-03  2.631e+00   0.002  0.99845
## green_rating              NA         NA      NA       NA
## net               -6.931e-01  1.111e+00  -0.624  0.53280
## amenities         -1.836e+00  6.416e-01  -2.862  0.00435 **
## cd_total_07       -1.815e-04  3.263e-04  -0.556  0.57824
## hd_total07         2.915e-04  2.249e-04   1.296  0.19553
## total_dd_07               NA         NA      NA       NA
## Precipitation      5.452e-02  3.979e-02   1.370  0.17106
## Gas_Costs         -3.533e+02  2.118e+02  -1.668  0.09581 .
## Electricity_Costs  1.318e+02  5.582e+01   2.361  0.01854 *
## cluster_rent       1.109e+00  3.303e-02  33.570  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.528 on 660 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.7515, Adjusted R-squared:  0.7447
## F-statistic: 110.9 on 18 and 660 DF,  p-value: < 2.2e-16
```

```
lm_green
```

```
##
## Call:
## lm(formula = Rent ~ cluster + CS_PropertyID + size + empl_gr +
##     leasing_rate + stories + age + renovated + class_a + class_b +
##     green_rating + net + amenities + cd_total_07 + hd_total07 +
##     total_dd_07 + Precipitation + Gas_Costs + Electricity_Costs +
##     cluster_rent, data = green)
##
## Coefficients:
```

```
##       (Intercept)            cluster      CS_PropertyID              size
##        -8.330e+00          1.597e-03          7.183e-07          5.554e-06
##           empl_gr       leasing_rate            stories               age
##         7.287e-02          3.679e-02         -2.799e-02          1.614e-02
##         renovated            class_a            class_b       green_rating
##        -3.142e-01          1.260e+00          5.121e-03                 NA
##               net          amenities        cd_total_07         hd_total07
##        -6.931e-01         -1.836e+00         -1.815e-04          2.915e-04
##       total_dd_07      Precipitation          Gas_Costs  Electricity_Costs
##                NA          5.452e-02         -3.533e+02          1.318e+02
##      cluster_rent
##         1.109e+00
```

```
lm_notgreen <- lm(Rent ~ cluster + CS_PropertyID + size + empl_gr + leasing_rate
          + stories + age + renovated + class_a + class_b + green_rating + net +
            amenities + cd_total_07 + hd_total07 + total_dd_07 +
            Precipitation + Gas_Costs + Electricity_Costs +
            cluster_rent, data = not_green)
summary(lm_notgreen)
```

```
##
## Call:
## lm(formula = Rent ~ cluster + CS_PropertyID + size + empl_gr +
##     leasing_rate + stories + age + renovated + class_a + class_b +
##     green_rating + net + amenities + cd_total_07 + hd_total07 +
##     total_dd_07 + Precipitation + Gas_Costs + Electricity_Costs +
##     cluster_rent, data = not_green)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.560  -3.646  -0.527   2.509 173.791
##
## Coefficients: (2 not defined because of singularities)
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -8.319e+00  1.075e+00  -7.738 1.15e-14 ***
## cluster            7.090e-04  3.037e-04   2.335 0.019572 *
## CS_PropertyID      2.869e-07  1.643e-07   1.746 0.080797 .
## size               6.853e-06  6.850e-07  10.003  < 2e-16 ***
## empl_gr            6.440e-02  1.856e-02   3.470 0.000523 ***
## leasing_rate       8.117e-03  5.560e-03   1.460 0.144360
## stories           -3.458e-02  1.720e-02  -2.010 0.044492 *
## age               -1.272e-02  4.936e-03  -2.577 0.009995 **
## renovated         -1.643e-01  2.750e-01  -0.597 0.550361
## class_a            2.867e+00  4.590e-01   6.247 4.41e-10 ***
## class_b            1.149e+00  3.533e-01   3.253 0.001149 **
## green_rating              NA         NA      NA       NA
## net               -2.741e+00  6.546e-01  -4.187 2.86e-05 ***
## amenities          8.227e-01  2.688e-01   3.060 0.002221 **
## cd_total_07       -1.382e-04  1.592e-04  -0.868 0.385296
## hd_total07         5.611e-04  9.630e-05   5.826 5.92e-09 ***
## total_dd_07               NA         NA      NA       NA
## Precipitation      5.072e-02  1.733e-02   2.926 0.003447 **
## Gas_Costs         -3.578e+02  8.382e+01  -4.269 1.99e-05 ***
## Electricity_Costs  1.949e+02  2.714e+01   7.182 7.57e-13 ***
```

```
## cluster_rent       9.998e-01  1.539e-02  64.949  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.633 on 7122 degrees of freedom
##   (68 observations deleted due to missingness)
## Multiple R-squared:  0.6039, Adjusted R-squared:  0.6029
## F-statistic: 603.4 on 18 and 7122 DF,  p-value: < 2.2e-16
```

```
lm_notgreen
```

```
##
## Call:
## lm(formula = Rent ~ cluster + CS_PropertyID + size + empl_gr +
##     leasing_rate + stories + age + renovated + class_a + class_b +
##     green_rating + net + amenities + cd_total_07 + hd_total07 +
##     total_dd_07 + Precipitation + Gas_Costs + Electricity_Costs +
##     cluster_rent, data = not_green)
##
## Coefficients:
##       (Intercept)            cluster      CS_PropertyID              size
##        -8.319e+00          7.090e-04          2.869e-07          6.853e-06
##           empl_gr       leasing_rate            stories                age
##         6.440e-02          8.117e-03         -3.458e-02         -1.272e-02
##         renovated            class_a            class_b       green_rating
##        -1.643e-01          2.867e+00          1.149e+00                 NA
##               net          amenities        cd_total_07         hd_total07
##        -2.741e+00          8.227e-01         -1.382e-04          5.611e-04
##       total_dd_07      Precipitation          Gas_Costs  Electricity_Costs
##                NA          5.072e-02         -3.578e+02          1.949e+02
##      cluster_rent
##         9.998e-01
```

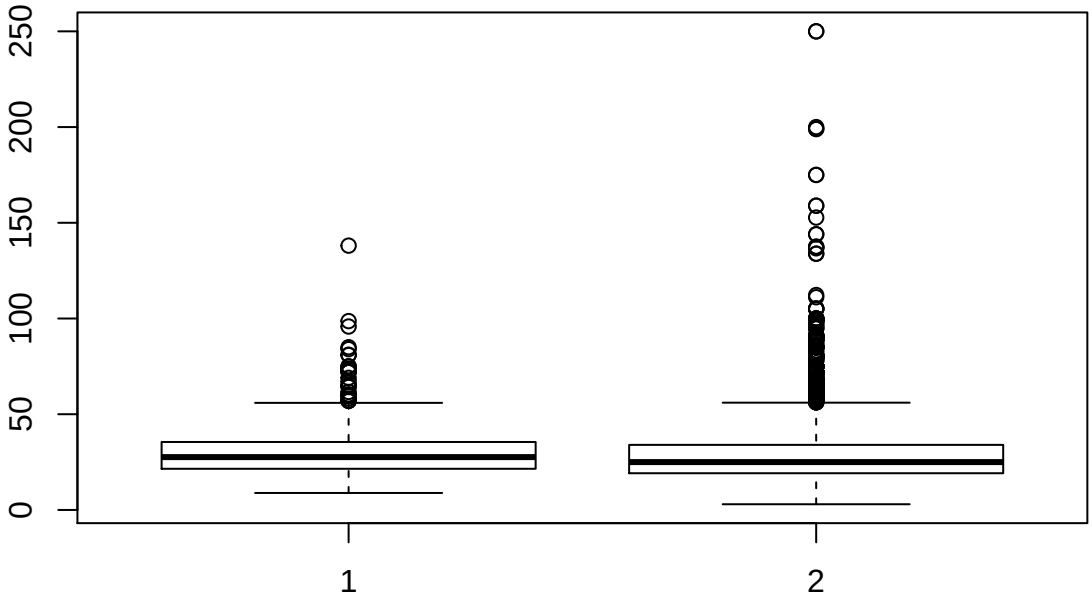
```
mean(green$Rent)
```

```
## [1] 30.01603
```

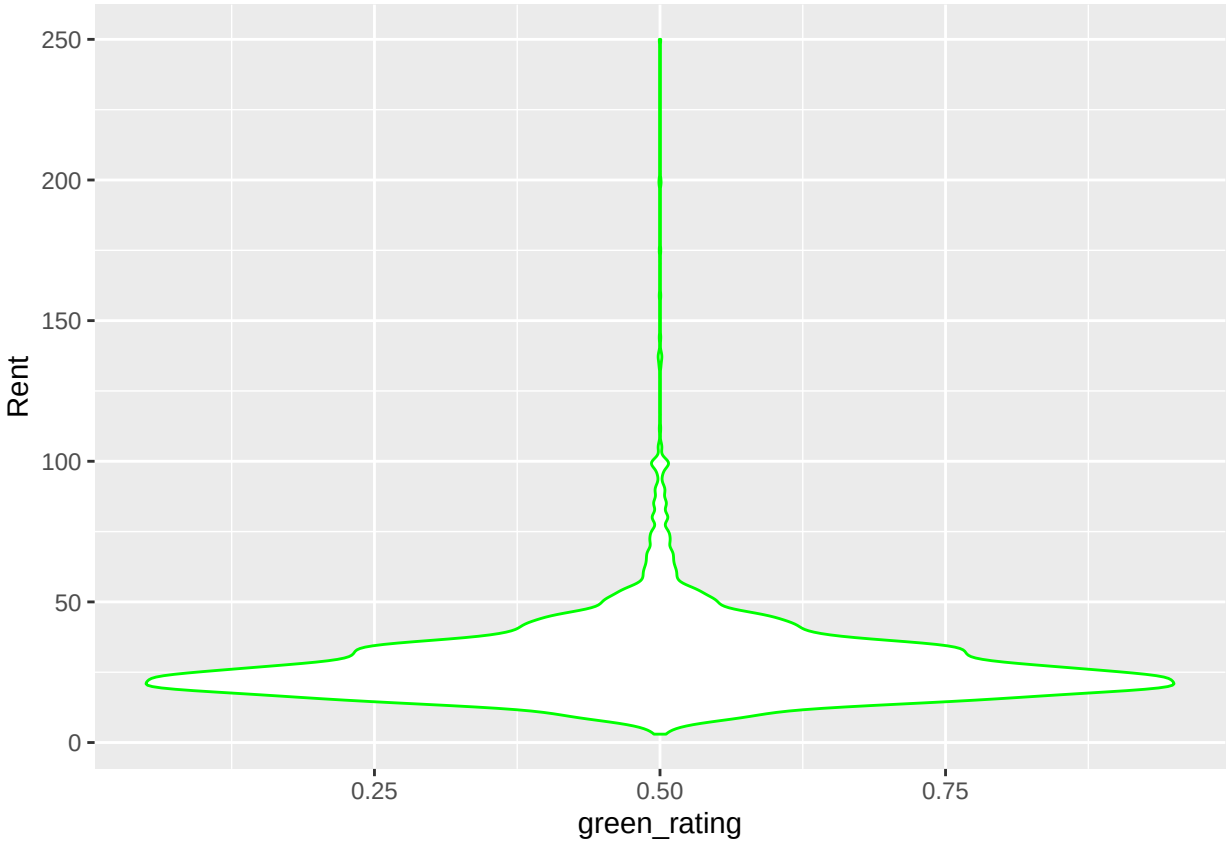
```
mean(not_green$Rent)
```

```
## [1] 28.26678
```

```
boxplot(green$Rent, not_green$Rent)
```




```
ggplot(data = greenbuildings)+
  geom_violin(mapping = aes(x=green_rating, y = Rent), color = 'green')
```




```
#The figures below picture the difference in rent prices with or without a green certification. Accordi

#Conclusion: While rent per square foot is higher on average on a green certified building, it is not a
```

```
#Problem 2:
#1. Running a regression on crime and police from different cities would just exemplify a correlation r
#2. The researchers at UPENN found data on crime in Washington DC by tracking alerts for potential terr
#3. They had to control METRO ridership because if people were not riding the metros, therefore traveli
#4. In this table, the researchers are showing crime in different locations and whether or not one loca
```

3) Wine Problem

Question

This data provides 11 chemical attributes of wine along with a quality rating and classification as white or red. We analyze the data of the chemical properties to first try to classify the wines as white or red and second to try to predict the quality ratings based on those chemical properties.

Method

We use both clustering and principal components analysis (PCA) to attempt to categorize the wines.

Clustering:

```
library(ggplot2)
library(LICORS)  # for kmeans++
library(foreach)
library(mosaic)
wine <- read.csv("~/GitHub/Class Folder/SDS323/data/wine.csv")

X = wine[,-(12:13)]
head(X)
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.4             0.70        0.00            1.9     0.076
## 2           7.8             0.88        0.00            2.6     0.098
## 3           7.8             0.76        0.04            2.3     0.092
## 4          11.2             0.28        0.56            1.9     0.075
## 5           7.4             0.70        0.00            1.9     0.076
## 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
## 2                  25                   67  0.9968 3.20      0.68     9.8
## 3                  15                   54  0.9970 3.26      0.65     9.8
## 4                  17                   60  0.9980 3.16      0.58     9.8
## 5                  11                   34  0.9978 3.51      0.56     9.4
## 6                  13                   40  0.9978 3.51      0.56     9.4
```

```
X = scale(X, center=TRUE, scale=TRUE)

mu = attr(X,"scaled:center")
sigma = attr(X,"scaled:scale")

clust2 = kmeanspp(X, k=2, nstart=25)

clust2$center[1,]*sigma + mu
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##            8.2895922            0.5319416            0.2695435
##       residual.sugar            chlorides  free.sulfur.dioxide
##            2.6342666            0.0883238           15.7647596
## total.sulfur.dioxide              density                   pH
##           48.6396835            0.9967404            3.3097200
##            sulphates              alcohol
##            0.6567194           10.4015216
```

```
clust2$center[2,]*sigma + mu
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##           6.85167903           0.27458385           0.33524928
##       residual.sugar            chlorides  free.sulfur.dioxide
##           6.39402555           0.04510424           35.52152864
## total.sulfur.dioxide              density                   pH
```

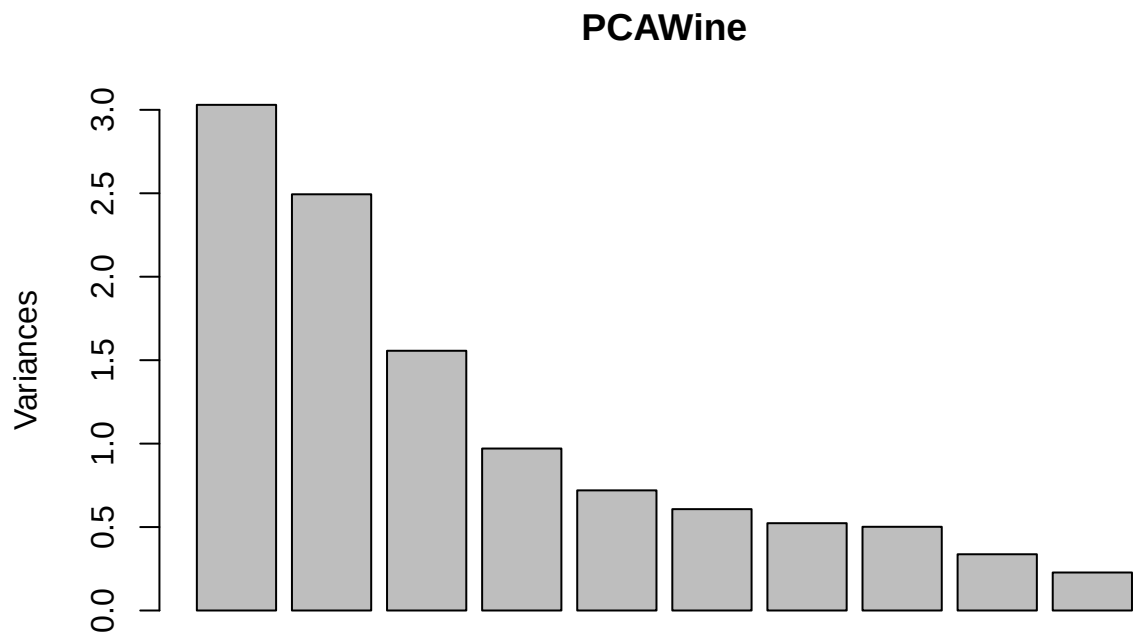
```
##          138.45848785             0.99400486             3.18762464
##             sulphates                alcohol
##             0.48880511            10.52235888
```

PCA:

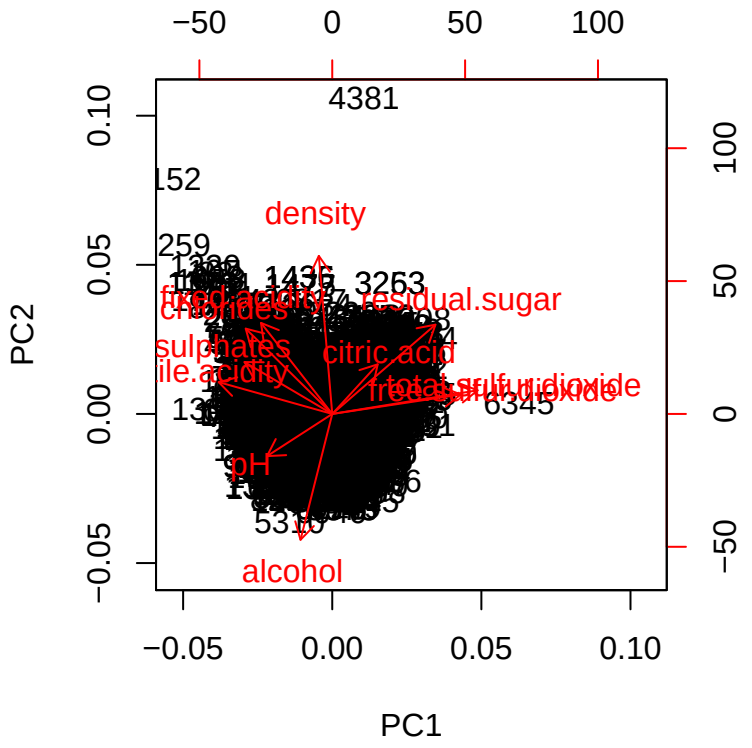
```
library(tidyverse)
library(ggplot2)
wine <- read.csv("~/GitHub/Class Folder/SDS323/data/wine.csv")

wine.type = wine$color
wine.quality = wine$quality

X = wine[,-(12:13)]
X = scale(X, center=TRUE, scale=TRUE)
PCAWine = prcomp(X, scale=TRUE)
plot(PCAWine)
```




```
biplot(PCAWine)
```




```
scores = PCAWine$x
```

Results

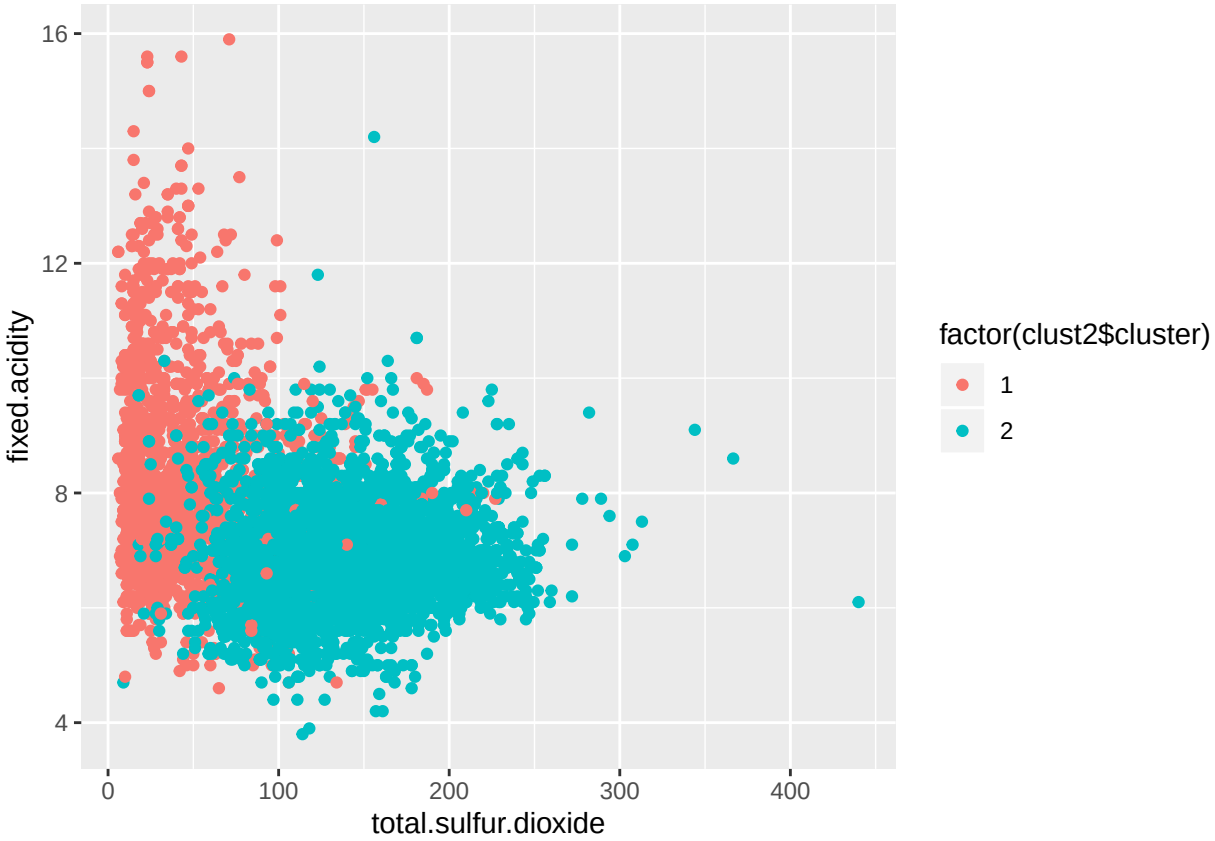
We compare the scatterplots generated by clustering and by PCA. We can also compare these scatterplots to the true classification by color according to the dataset.

For the cluster and original scatterplots, we use total sulfur dioxide and fixed acidity for the x and y axis becayse these variables most clearly show the variations in wine. Likewise, for the PCA plots, we use components 1 and 2 because they show the data most clearly.

Color Scatterplots:

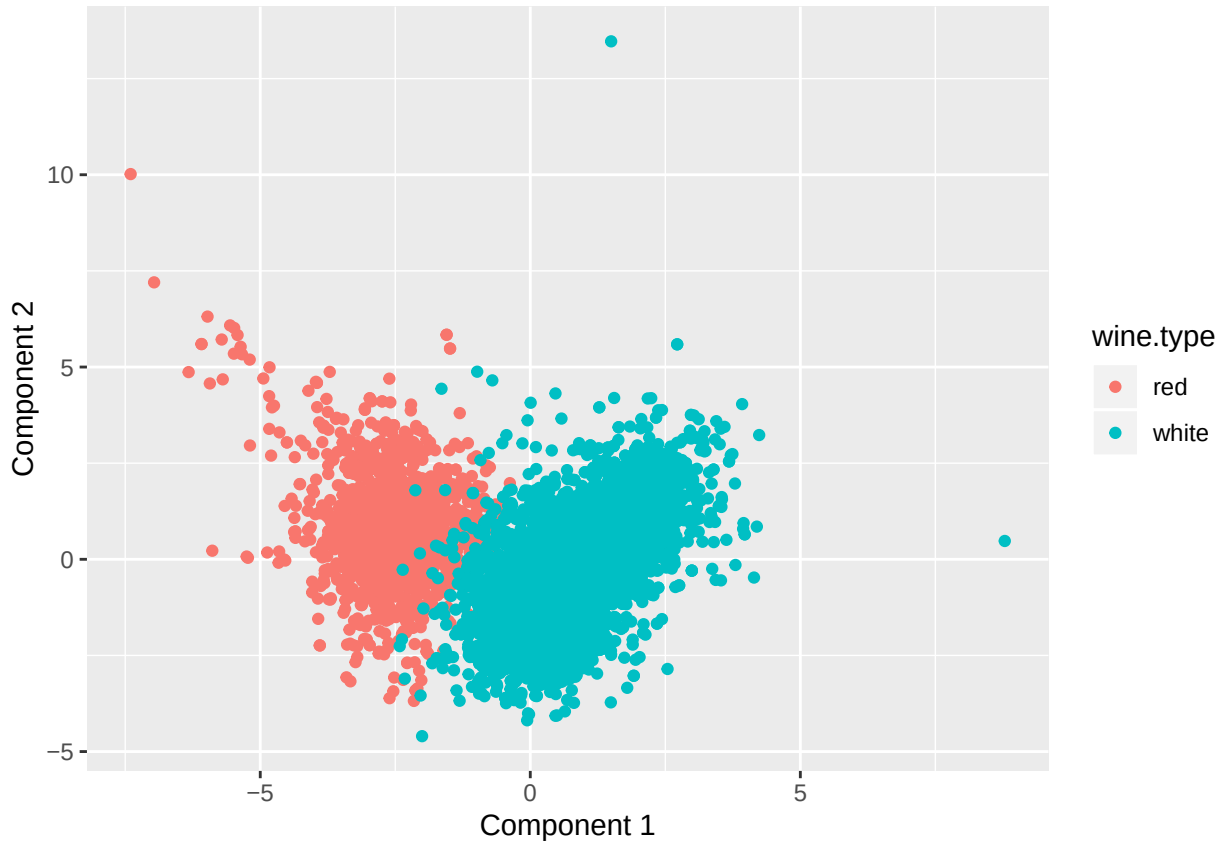
Clustering:

```
qplot(total.sulfur.dioxide, fixed.acidity, data=wine, color=factor(clust2$cluster))
```



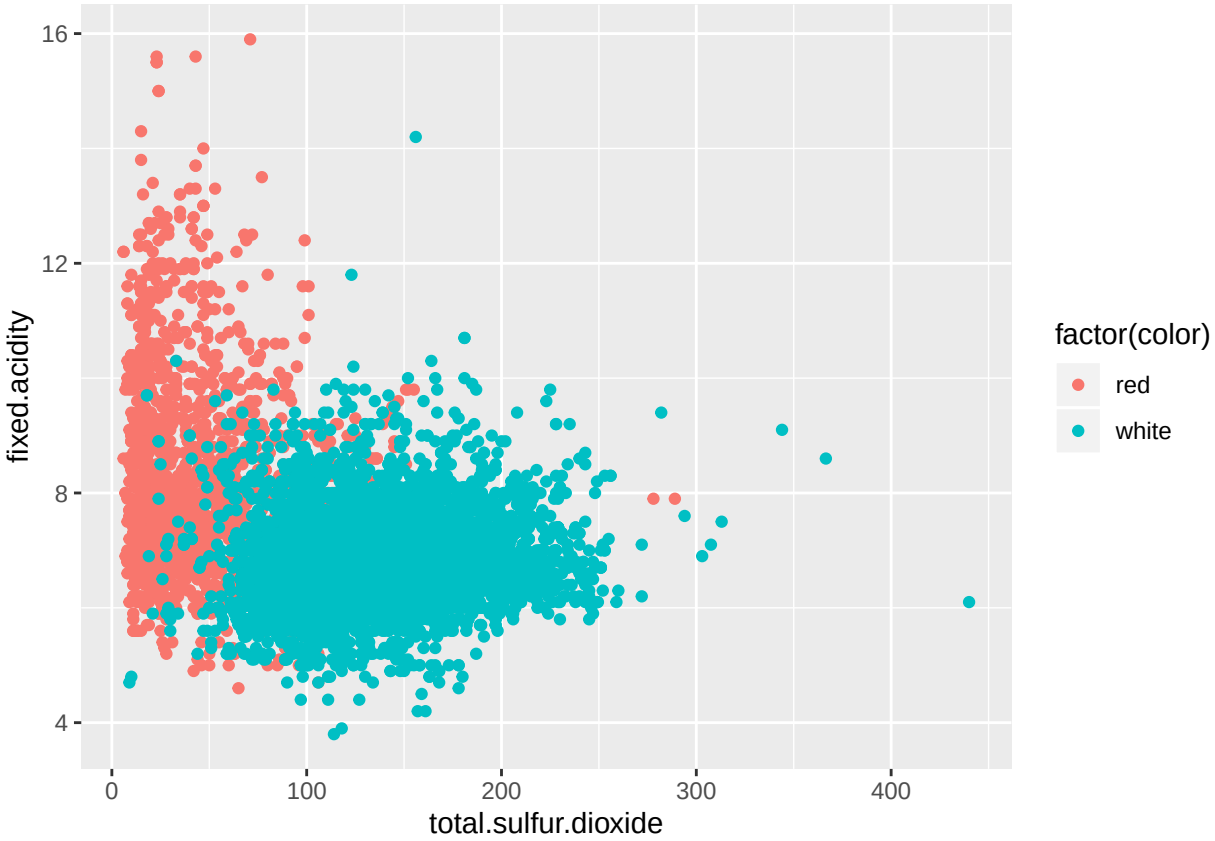
PCA:

```
qplot(scores[,1], scores[,2], color=wine.type, xlab='Component 1', ylab='Component 2')
```



Original:

```
qplot(total.sulfur.dioxide, fixed.acidity,  data=wine, color=factor(color))
```

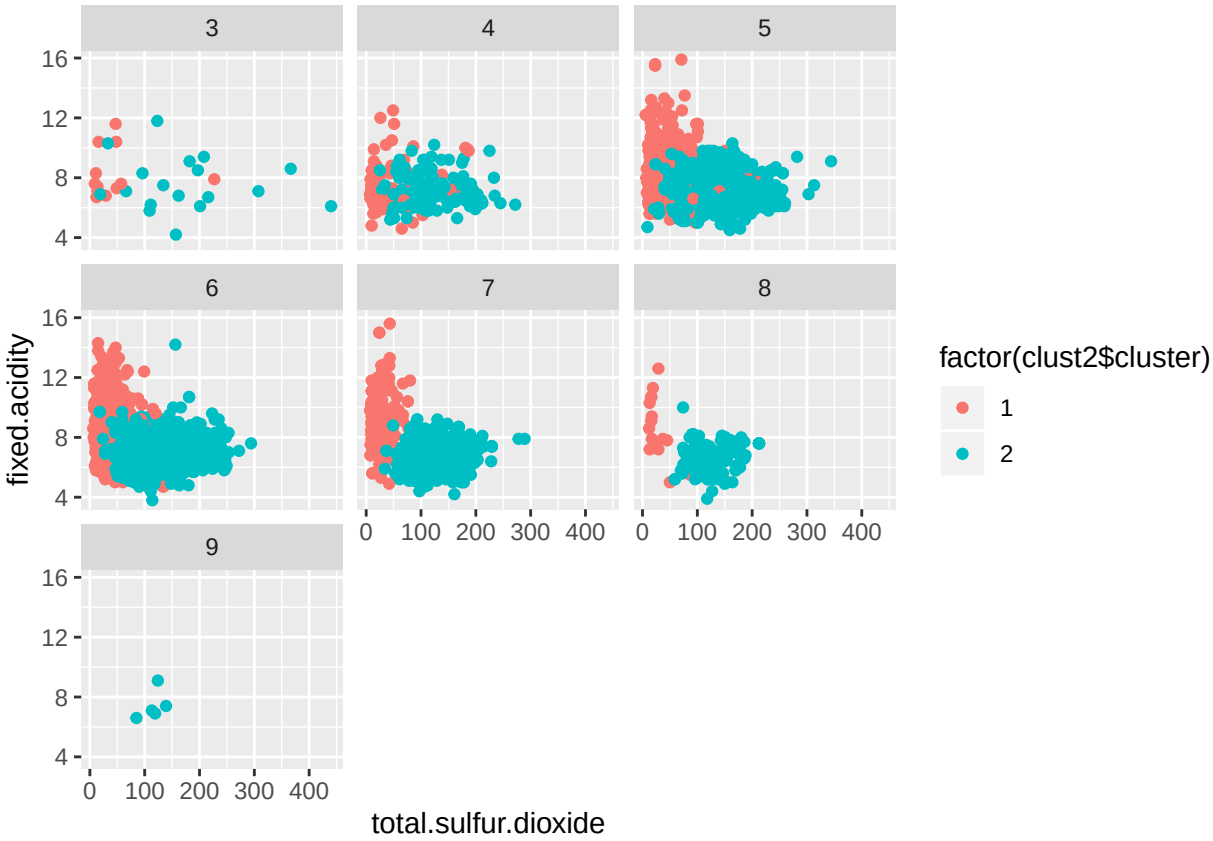


Quality Scatterplots

We also try to identify patterns between the chemical make-up of the wines and their quality ratings. Because the wines are rated in whole numbers between 1 and 10, we can put these plots side-by-side to look for differences between each level of quality.

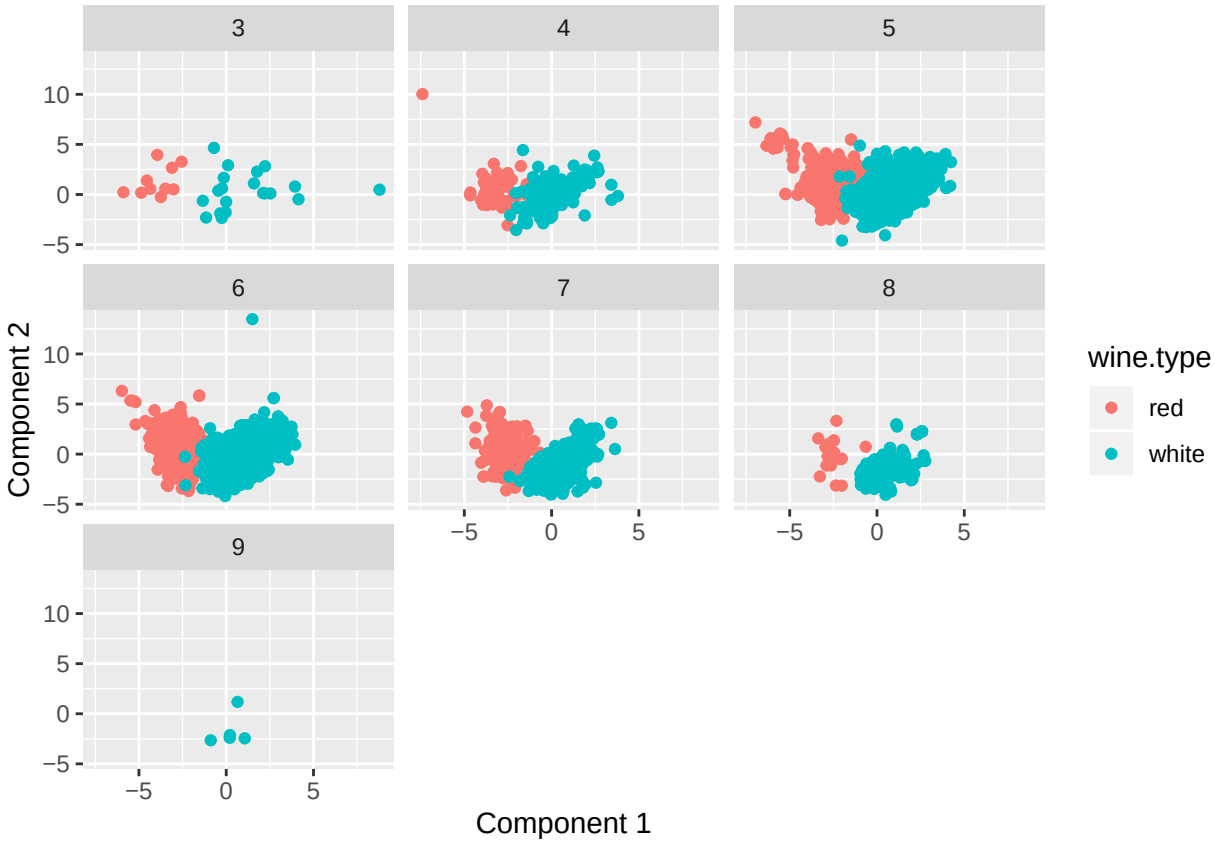
Clustering:

```
qplot(total.sulfur.dioxide, fixed.acidity, data=wine, facets=~wine$quality, color=factor(clust2$cluster
```



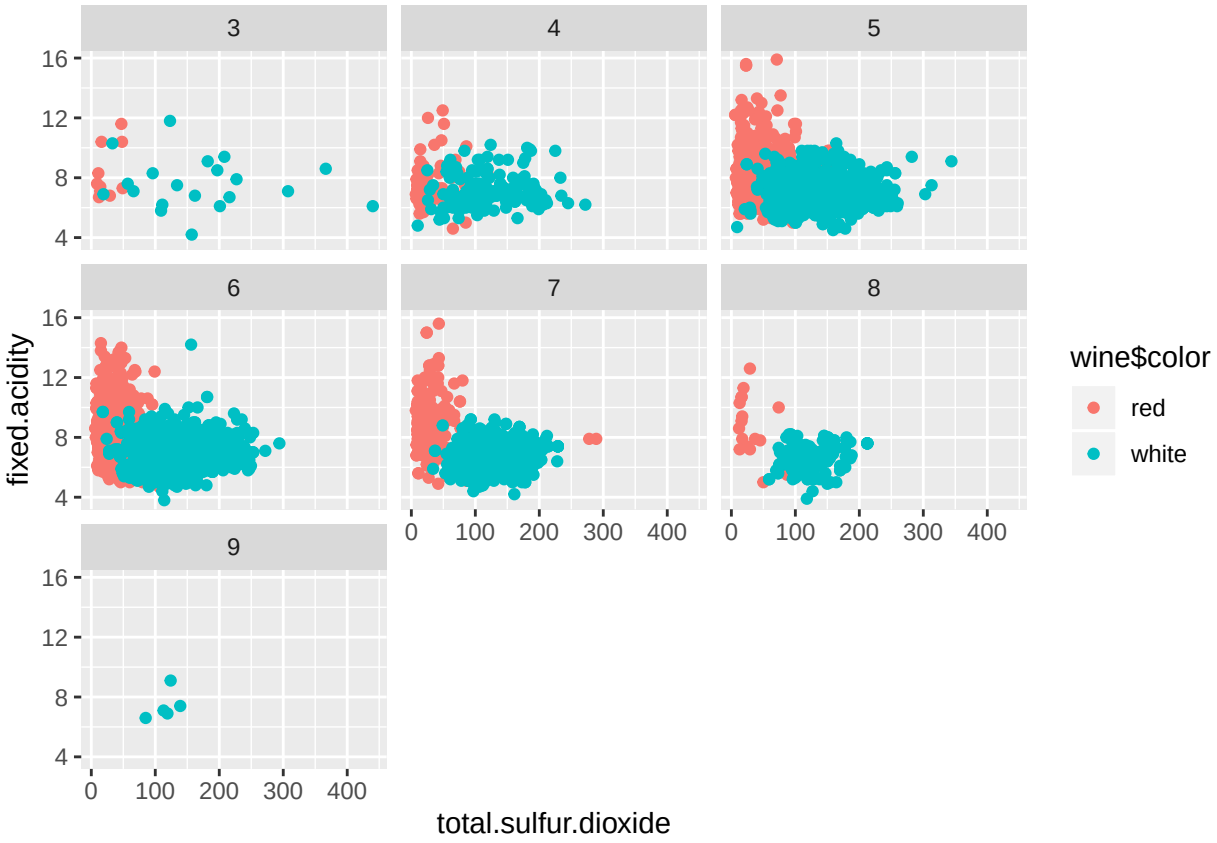
PCA:

```
qplot(scores[,1], scores[,2], facets=~wine.quality, color=wine.type, xlab='Component 1', ylab='Componen
```



Original:

```
qplot(total.sulfur.dioxide, fixed.acidity, facets=~wine$quality, data=wine, color=wine$color)
```



Results

For categorizing wines as red or white, principal components analysis (PCA) is better at clearly classifying wines. The scatterplots for clustering show considerable overlap for red and white wines. This makes sense because wines share many chemical properties regardless of their different colors. Therefore, there are overlapping factors in both red and white wines. Clustering based upon chemical properties would be insufficient because clusters must be mutually exclusive. PCA identifies the most meaningful differentiating properties while also tolerating overlapping attributes.

However, both clustering and PCA are insufficient to categorize wines by quality rating based on chemical identifiers. Wines seem to have a similar spread of chemical properties at every rating level. Thus, wine quality seems to be driven by non-chemical factors or chemical factors that are too subtle to be detected through either method of categorization.

4) Social Marketing

Question

How can social media be used to identify market segments for a product?

Method

We standardize the data before running the clustering algorithims. This enables us to identify groups of customers based on common interest. These interests can later be used to tailor marketing approaches.

```
library(ggplot2)
library(LICORS)
library(foreach)
library(mosaic)

tweets=read.csv("~/GitHub/Class Folder/SDS323/data/social_marketing.csv")
summary(tweets)
```

```
##          X           chatter       current_events      travel
##  123pxkyqj:   1   Min.   : 0.000   Min.   :0.000   Min.   : 0.000
##  12grikctu:   1   1st Qu.: 2.000   1st Qu.:1.000   1st Qu.: 0.000
##  12klxic7j:   1   Median : 3.000   Median :1.000   Median : 1.000
##  12t4msroj:   1   Mean   : 4.399   Mean   :1.526   Mean   : 1.585
##  12yam59l3:   1   3rd Qu.: 6.000   3rd Qu.:2.000   3rd Qu.: 2.000
##  132y8f6aj:   1   Max.   :26.000   Max.   :8.000   Max.   :26.000
##  (Other)  :7876
##  photo_sharing    uncategorized      tv_film      sports_fandom
##  Min.   : 0.000   Min.   :0.000   Min.   : 0.00   Min.   : 0.000
##  1st Qu.: 1.000   1st Qu.:0.000   1st Qu.: 0.00   1st Qu.: 0.000
##  Median : 2.000   Median :1.000   Median : 1.00   Median : 1.000
##  Mean   : 2.697   Mean   :0.813   Mean   : 1.07   Mean   : 1.594
##  3rd Qu.: 4.000   3rd Qu.:1.000   3rd Qu.: 1.00   3rd Qu.: 2.000
##  Max.   :21.000   Max.   :9.000   Max.   :17.00   Max.   :20.000
##
##     politics           food            family        home_and_garden
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.0000   Min.   :0.0000
##  1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.0000   1st Qu.:0.0000
##  Median : 1.000   Median : 1.000   Median : 1.0000   Median :0.0000
##  Mean   : 1.789   Mean   : 1.397   Mean   : 0.8639   Mean   :0.5207
##  3rd Qu.: 2.000   3rd Qu.: 2.000   3rd Qu.: 1.0000   3rd Qu.:1.0000
##  Max.   :37.000   Max.   :16.000   Max.   :10.0000   Max.   :5.0000
##
##      music              news        online_gaming       shopping
##  Min.   : 0.0000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 0.0000   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000
##  Median : 0.0000   Median : 0.000   Median : 0.000   Median : 1.000
##  Mean   : 0.6793   Mean   : 1.206   Mean   : 1.209   Mean   : 1.389
##  3rd Qu.: 1.0000   3rd Qu.: 1.000   3rd Qu.: 1.000   3rd Qu.: 2.000
##  Max.   :13.0000   Max.   :20.000   Max.   :27.000   Max.   :12.000
##
##  health_nutrition  college_uni     sports_playing      cooking
##  Min.   : 0.000   Min.   : 0.000   Min.   :0.0000   Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.: 0.000
##  Median : 1.000   Median : 1.000   Median :0.0000   Median : 1.000
##  Mean   : 2.567   Mean   : 1.549   Mean   :0.6392   Mean   : 1.998
##  3rd Qu.: 3.000   3rd Qu.: 2.000   3rd Qu.:1.0000   3rd Qu.: 2.000
##  Max.   :41.000   Max.   :30.000   Max.   :8.0000   Max.   :33.000
##
##       eco           computers          business         outdoors
```

```
##  Min.   :0.0000   Min.   : 0.0000   Min.   :0.0000   Min.   : 0.0000
##  1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.: 0.0000
##  Median :0.0000   Median : 0.0000   Median :0.0000   Median : 0.0000
##  Mean   :0.5123   Mean   : 0.6491   Mean   :0.4232   Mean   : 0.7827
##  3rd Qu.:1.0000   3rd Qu.: 1.0000   3rd Qu.:1.0000   3rd Qu.: 1.0000
##  Max.   :6.0000   Max.   :16.0000   Max.   :6.0000   Max.   :12.0000
##
##      crafts         automotive          art             religion
##  Min.   :0.0000   Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.000
##  1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.000
##  Median :0.0000   Median : 0.0000   Median : 0.0000   Median : 0.000
##  Mean   :0.5159   Mean   : 0.8299   Mean   : 0.7248   Mean   : 1.095
##  3rd Qu.:1.0000   3rd Qu.: 1.0000   3rd Qu.: 1.0000   3rd Qu.: 1.000
##  Max.   :7.0000   Max.   :13.0000   Max.   :18.0000   Max.   :20.000
##
##      beauty          parenting          dating            school
##  Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000
##  1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000
##  Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 0.0000
##  Mean   : 0.7052   Mean   : 0.9213   Mean   : 0.7109   Mean   : 0.7677
##  3rd Qu.: 1.0000   3rd Qu.: 1.0000   3rd Qu.: 1.0000   3rd Qu.: 1.0000
##  Max.   :14.0000   Max.   :14.0000   Max.   :24.0000   Max.   :11.0000
##
##  personal_fitness    fashion        small_business        spam
##  Min.   : 0.000   Min.   : 0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.: 0.000   1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median : 0.000   Median : 0.0000   Median :0.0000   Median :0.00000
##  Mean   : 1.462   Mean   : 0.9966   Mean   :0.3363   Mean   :0.00647
##  3rd Qu.: 2.000   3rd Qu.: 1.0000   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :19.000   Max.   :18.0000   Max.   :6.0000   Max.   :2.00000
##
##      adult
##  Min.   : 0.0000
##  1st Qu.: 0.0000
##  Median : 0.0000
##  Mean   : 0.4033
##  3rd Qu.: 0.0000
##  Max.   :26.0000
##
```

```
x=tweets[,-(1:1)]
Z = scale(x, center=TRUE, scale=TRUE)
head(Z)
```

```
##         chatter current_events     travel photo_sharing uncategorized
## [1,] -0.6797028     -1.2028323  0.1815755    -0.2550887     1.2683700
## [2,] -0.3963465      1.1614380  0.1815755    -0.6211866     0.1998266
## [3,]  0.4537224      1.1614380  1.0566463     0.1110091     0.1998266
## [4,] -0.9630591      2.7376183  0.1815755    -0.2550887    -0.8687169
## [5,]  0.1703661      0.3733479 -0.6934952     1.2093027     0.1998266
## [6,]  0.4537224      1.9495282  0.1815755     1.5754006    -0.8687169
##          tv_film sports_fandom    politics       food    family home_and_garden
## [1,] -0.04237246    -0.2748886 -0.59009085  1.4657438 0.1201991       2.0080596
## [2,] -0.04237246     1.1134105 -0.26017908  0.3393369 1.0031532       0.6506389
```

```
## [3,]  2.36903350    -0.7376550  0.06973268 -0.2238666 0.1201991       0.6506389
## [4,] -0.04237246    -0.7376550 -0.26017908 -0.7870700 0.1201991      -0.7067819
## [5,] -0.64522395    -0.7376550  0.06973268 -0.7870700 0.1201991      -0.7067819
## [6,] -0.04237246    -0.2748886 -0.59009085  0.3393369 0.1201991       0.6506389
##           music        news online_gaming   shopping health_nutrition
## [1,] -0.6594752 -0.57384947    -0.4498032 -0.2152578        3.2100303
## [2,] -0.6594752 -0.57384947    -0.4498032 -0.7680965       -0.5709875
## [3,]  0.3113846 -0.09783584    -0.4498032  0.3375809       -0.5709875
## [4,] -0.6594752 -0.57384947    -0.4498032 -0.7680965       -0.5709875
## [5,] -0.6594752 -0.57384947     0.6664905  0.3375809       -0.5709875
## [6,]  0.3113846 -0.57384947    -0.4498032  1.9960970       -0.5709875
##      college_uni sports_playing       cooking        eco  computers   business
## [1,]  -0.5348282      1.3949754  0.8751685215  0.6335945  0.2975329 -0.6112878
## [2,]  -0.5348282      0.3698779 -0.5825825957 -0.6655710 -0.5503175  0.8330079
## [3,]  -0.5348282     -0.6552197  0.0005178512  0.6335945 -0.5503175 -0.6112878
## [4,]  -0.1896619     -0.6552197 -0.5825825957 -0.6655710 -0.5503175  0.8330079
## [5,]   0.8458369     -0.6552197 -0.2910323722 -0.6655710  0.2975329 -0.6112878
## [6,]  -0.5348282     -0.6552197 -0.5825825957 -0.6655710  0.2975329  0.8330079
##        outdoors     crafts automotive        art    religion     beauty
## [1,]  1.0064885  0.5926945 -0.6074479 -0.4447882 -0.04982546 -0.5310261
## [2,] -0.6471107  1.8169132 -0.6074479 -0.4447882 -0.57206517 -0.5310261
## [3,] -0.6471107  1.8169132 -0.6074479  4.4644655 -0.57206517  0.2220411
## [4,] -0.6471107  3.0411319 -0.6074479  0.7825252 -0.57206517  0.2220411
## [5,]  0.1796889 -0.6315241 -0.6074479 -0.4447882 -0.57206517 -0.5310261
## [6,] -0.6471107 -0.6315241  0.1245356 -0.4447882 -0.57206517 -0.5310261
##       parenting     dating     school personal_fitness      fashion
## [1,]  0.05190867  0.1622242 -0.6460699        3.9654745 -0.545049065
## [2,] -0.60800117  0.1622242  2.7201990       -0.6078657 -0.545049065
## [3,] -0.60800117  0.1622242 -0.6460699       -0.6078657  0.001873498
## [4,] -0.60800117 -0.3988338 -0.6460699       -0.6078657 -0.545049065
## [5,] -0.60800117 -0.3988338 -0.6460699       -0.6078657 -0.545049065
## [6,] -0.60800117 -0.3988338 -0.6460699       -0.6078657 -0.545049065
##      small_business        spam      adult
## [1,]     -0.5441037 -0.07768727 -0.2224097
## [2,]     -0.5441037 -0.07768727 -0.2224097
## [3,]     -0.5441037 -0.07768727 -0.2224097
## [4,]     -0.5441037 -0.07768727 -0.2224097
## [5,]      1.0736350 -0.07768727 -0.2224097
## [6,]     -0.5441037 -0.07768727 -0.2224097
```

```
mu = attr(Z,"scaled:center")
sigma = attr(Z,"scaled:scale")

clust = kmeanspp(Z, k=4, nstart=25)

clust$center[1,]*sigma + mu
```

```
##          chatter   current_events           travel    photo_sharing
##      4.404761905      1.656862745      5.627450980      2.445378151
##    uncategorized          tv_film    sports_fandom         politics
##      0.782913165      1.142857143      2.042016807      8.990196078
##             food           family  home_and_garden            music
##      1.460784314      0.929971989      0.610644258      0.633053221
##             news    online_gaming         shopping health_nutrition
```

```
##      5.284313725      1.138655462      1.301120448      2.029411765
##      college_uni   sports_playing          cooking              eco
##      1.532212885      0.707282913      1.406162465      0.591036415
##        computers         business         outdoors           crafts
##      2.476190476      0.644257703      1.001400560      0.607843137
##       automotive              art         religion           beauty
##      2.362745098      0.679271709      1.023809524      0.512605042
##        parenting           dating           school personal_fitness
##      0.960784314      1.047619048      0.722689076      1.189075630
##          fashion   small_business             spam            adult
##      0.731092437      0.473389356      0.008403361      0.238095238
```

```
clust$center[2,]*sigma + mu
```

```
##          chatter   current_events           travel    photo_sharing
##      4.109375000      1.679687500      1.342447917      2.548177083
##    uncategorized          tv_film    sports_fandom         politics
##      0.746093750      1.052083333      5.962239583      1.186197917
##             food           family  home_and_garden            music
##      4.609375000      2.519531250      0.648437500      0.726562500
##             news    online_gaming         shopping health_nutrition
##      1.039062500      1.272135417      1.404947917      2.182291667
##      college_uni   sports_playing          cooking              eco
##      1.454427083      0.766927083      1.733072917      0.652343750
##        computers         business         outdoors           crafts
##      0.743489583      0.503906250      0.748697917      1.080729167
##       automotive              art         religion           beauty
##      1.050781250      0.884114583      5.364583333      1.106770833
##        parenting           dating           school personal_fitness
##      4.104166667      0.664062500      2.704427083      1.394531250
##          fashion   small_business             spam            adult
##      1.040364583      0.389322917      0.006510417      0.425781250
```

```
clust$center[3,]*sigma + mu
```

```
##          chatter   current_events           travel    photo_sharing
##      3.666229508      1.371366120      1.064043716      1.883715847
##    uncategorized          tv_film    sports_fandom         politics
##      0.672349727      0.825573770      0.943387978      0.954535519
##             food           family  home_and_garden            music
##      0.800000000      0.556939891      0.406120219      0.478032787
##             news    online_gaming         shopping health_nutrition
##      0.692240437      0.935737705      0.979890710      1.528961749
##      college_uni   sports_playing          cooking              eco
##      1.155191257      0.451803279      0.931147541      0.347759563
##        computers         business         outdoors           crafts
##      0.356284153      0.289617486      0.490710383      0.324153005
##       automotive              art         religion           beauty
##      0.542295082      0.493333333      0.516065574      0.320218579
##        parenting           dating           school personal_fitness
##      0.449180328      0.429289617      0.401311475      0.866229508
##          fashion   small_business             spam            adult
##      0.468415301      0.233005464      0.005464481      0.380109290
```

```
clust$center[4,]*sigma + mu
```

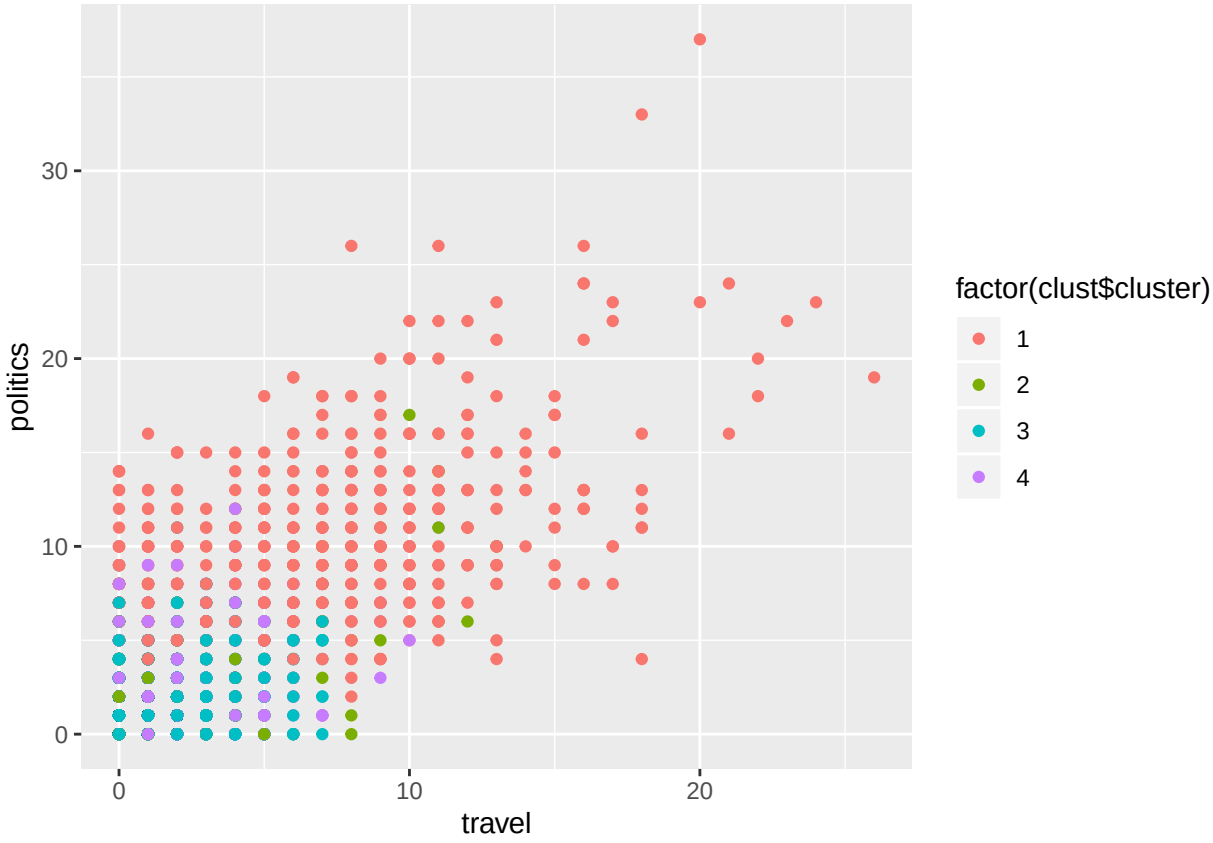
```
##          chatter   current_events           travel    photo_sharing
##      6.354520548      1.798904110      1.411506849      4.895890411
##    uncategorized          tv_film    sports_fandom         politics
##      1.205479452      1.663013699      1.211506849      1.315616438
##             food           family  home_and_garden            music
##      1.518904110      0.910684932      0.718904110      1.181917808
##             news    online_gaming         shopping health_nutrition
##      0.966575342      1.894246575      2.443835616      5.542465753
##      college_uni   sports_playing          cooking              eco
##      2.584657534      1.028493151      5.016438356      0.835068493
##        computers         business         outdoors           crafts
##      0.628493151      0.637808219      1.443287671      0.722739726
##       automotive              art         religion           beauty
##      0.858082192      1.255890411      0.779178082      1.576438356
##        parenting           dating           school personal_fitness
##      0.750136986      1.304657534      0.888767123      3.090958904
##          fashion   small_business             spam            adult
##      2.406027397      0.519452055      0.008219178      0.516712329
```

Results

Based upon the key factors in each cluster, we construct scatterplots that identify intersecting interests. Here are three scatterplots that identify possible market segments:

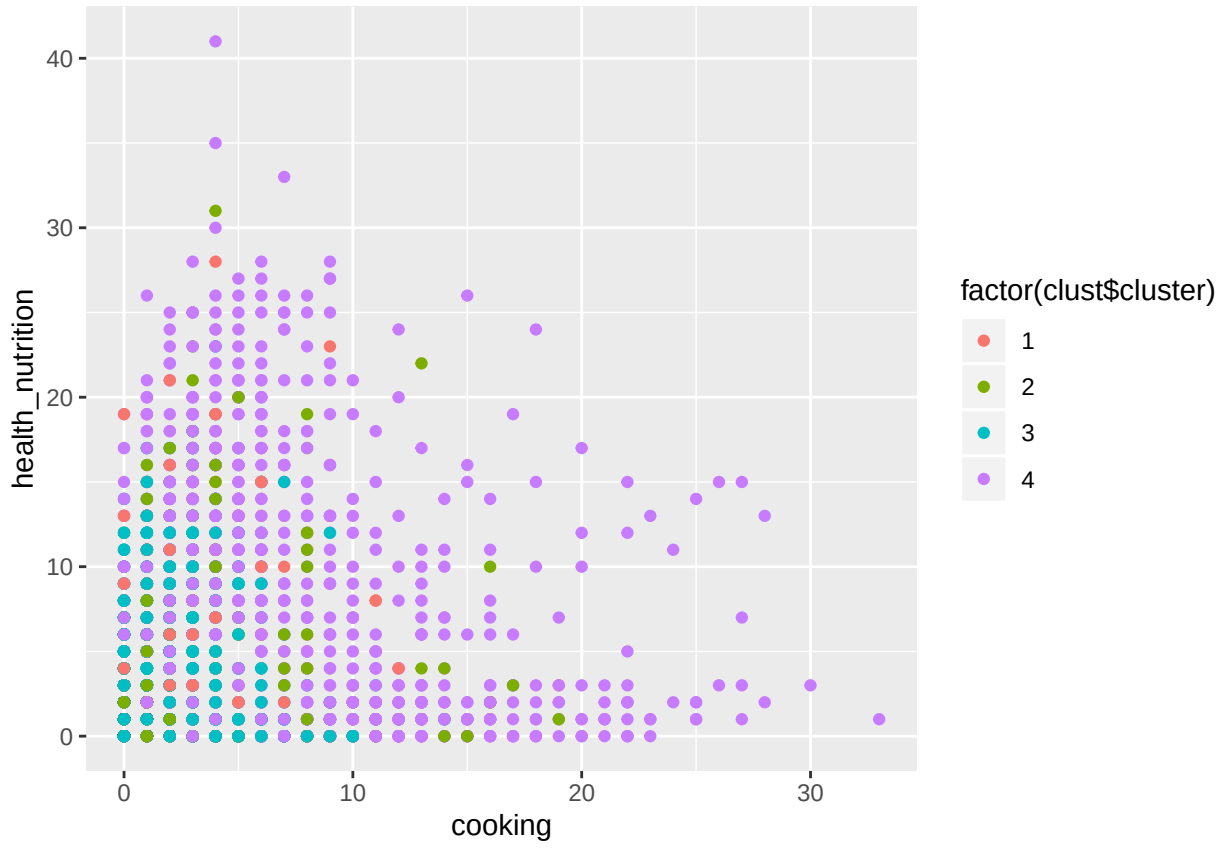
Travel and Politics

```
qplot(travel, politics, data=tweets, color=factor(clust$cluster))
```



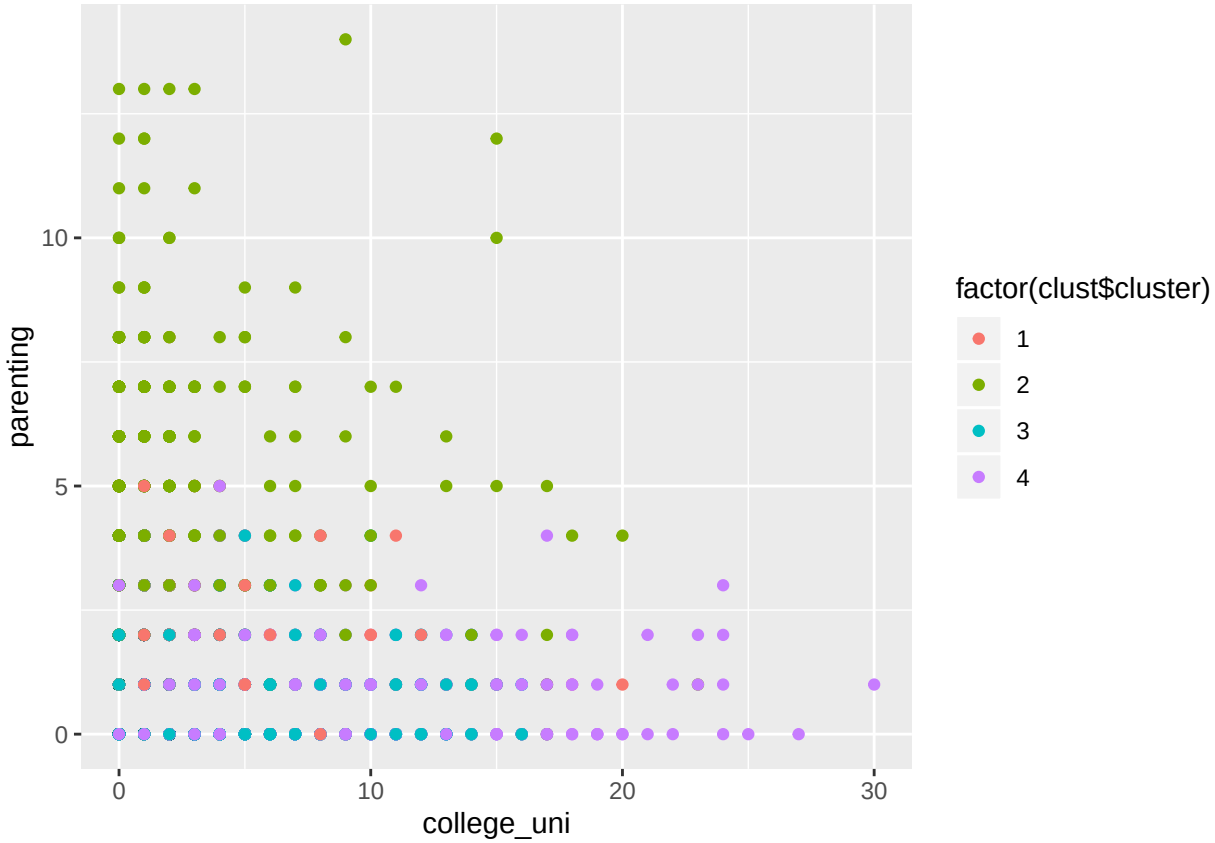
Health_nutrition and Cooking

```
qplot(cooking, health_nutrition, data=tweets, color=factor(clust$cluster))
```



Parenting and College_Uni

```
qplot(college_uni, parenting, data=tweets, color=factor(clust$cluster))
```



Conclusions

The first two graphs above are used to show co-relation, i.e. a significant cluster in the upper right quadrant of the plot shows that people in that cluster are highly interest in both topics. In the third, graph, we see two distinct clusters in the bottom right and upper left quadrants. Here are the ramifications of each graph for marketing strategies:

Travel and Politics: There is a high concentration of people tweeting about both travel and politics. This might be an indicator of the common socioeconomic status of consumers of this product. It might also point to particular vocations that require lots of travel for work and have a connection to politics, such as journalists or businesspeople. This information may justify placing ads in the online and print editions of political news sources, such as the New York Times. Moreover, the compan might also consider having the product sold in airports or at newsstands.

Health_Nutrition and Cooking: It may be unsurprising that people interest in health and nutrition are consumers of a nutrition supplement. However, the relationship with cooking may also point to possible marketing strategies. The company could send samples to cooking bloggers who would help promote the product. Also, selling this product in high-end grocery stores (like Central Market or Whole Foods) may be successful, because its shoppers are likely both interest in nutrition and cooking.

Parenting and College_Uni: This graph shows two distinct segments that might help understand the general age of people who consume this product. There is one significant cluster among people who tweet often about parenting but seldom about college. This might lead us to conclude that consumers of this product tend to be older, since there is no consistent pattern among people who tweet about college. This inference of age can help the company decide where to sell its products and where to post advertisements.