R notes:

* Feature scaling - sets mean to 0 and std to 1:

scale(x) - where x can be a matrix or vector or data frame

* Select column subset - If columns are labeled, use subset function:

dat <- data.frame(A=c(1,2),B=c(3,4),C=c(5,6),D=c(7,7),E=c(8,8),F=c(9,9))

subset(dat, select=c("A", "B"))

* Evaluate normality of features: Plot a Q-Q plot

qqnorm(z)

qqline(z)

* Leverages and outliers: Cook’s distance is a formula applied to find the outliers. Values > 1 are considered large

lm.influence(model)

* Plotting the model itself will give you 4 graphs:  Residuals vs. fitted, Normal Q-Q plot, Scale-location (standardized residuals), Residuals vs. leverages - with Cook’s distance lines
* Model Confidence: Find the confidence interval of each feature using:

confint(model, level=0.95)

Plot of fitted vs residuals. No clear pattern should show in the residual plot if the model is a good fit

Plot(fitted(model),residuals(model))

* Correlations: Compute correlation matrix of features

install.packages(“Car”); Library(car)

fit = …

vif(fit) 🡪 variance inflation factor – fraction of variance unexaplined by Model

|  |  |
| --- | --- |
| VIF = 1 | Not correlated |
| 1 < VIF < 5 | Moderately correlated |
| VIF > 5 to 10 | Highly correlated |

scatterplotMatrix(~temp + density + rate + am +defect, quality) 🡪 Gives a matrix of scatterplot, each variable against the other

* Preprocessing stuff:
  + Fine the average by factor (‘Zones’)

> library(plyr)

> r2<-ddply(data, .(Zone), summarize, mean=mean(Zone\_Section\_Average\_Price))

This gives a vector – r2, of mean by zone

Sort this with:

install.packages("data.table")

library(data.table)

sorted = data.table(r2, key="mean")

* After plotting the zone means I found a hockey stick shaped curve, where the means began to grow exponentially (after linear mean growth). The cutoff seemed to be at around $400. I will now split up the data sets to create 2 models, including the zones before and after $400

Exploratory Plan:

* Feature Transformation:
  + Find the necessary transformations by applying a scatterplot to each feature regressed on multiple estimators (different future days)
  + Normal QQ plot to test normality of different transformations

