Assignment 4

Shubham Maheshwari ID- 8894405

2024-11-18

# 1 Preliminary and Exploratory

df <- read.table("PROG8435-24F-Assign04.txt", sep = ",", header = TRUE )  
  
df <- as.data.frame(df)  
head(df)

## Prc Bed floor TotFloor Bath Sqft City Comp Dist  
## 1 596 3 3 4 4 501 Blossomville Leaseflow 10.6  
## 2 3130 2 5 7 2 1275 Terranova Rentopia 1.4  
## 3 2300 1 4 6 2 982 Riverport Leaseflow 6.8  
## 4 2840 1 2 4 1 2418 Terranova Leaseflow 0.4  
## 5 2790 4 1 2 3 1655 Terranova Rentopia 5.3  
## 6 3230 3 5 5 2 1024 Riverport Leaseflow 0.5

# Renaming all vairables with "SM"   
colnames(df) <- paste(colnames(df), "SM", sep = "\_")  
head(df)

## Prc\_SM Bed\_SM floor\_SM TotFloor\_SM Bath\_SM Sqft\_SM City\_SM Comp\_SM  
## 1 596 3 3 4 4 501 Blossomville Leaseflow  
## 2 3130 2 5 7 2 1275 Terranova Rentopia  
## 3 2300 1 4 6 2 982 Riverport Leaseflow  
## 4 2840 1 2 4 1 2418 Terranova Leaseflow  
## 5 2790 4 1 2 3 1655 Terranova Rentopia  
## 6 3230 3 5 5 2 1024 Riverport Leaseflow  
## Dist\_SM  
## 1 10.6  
## 2 1.4  
## 3 6.8  
## 4 0.4  
## 5 5.3  
## 6 0.5

cat("\nStructure of the Dataframe:\n")

##   
## Structure of the Dataframe:

str(df)

## 'data.frame': 1051 obs. of 9 variables:  
## $ Prc\_SM : num 596 3130 2300 2840 2790 3230 2240 2030 1800 1600 ...  
## $ Bed\_SM : int 3 2 1 1 4 3 1 3 4 1 ...  
## $ floor\_SM : int 3 5 4 2 1 5 1 2 1 1 ...  
## $ TotFloor\_SM: int 4 7 6 4 2 5 4 4 5 4 ...  
## $ Bath\_SM : int 4 2 2 1 3 2 1 1 1 1 ...  
## $ Sqft\_SM : int 501 1275 982 2418 1655 1024 864 671 609 834 ...  
## $ City\_SM : chr "Blossomville" "Terranova" "Riverport" "Terranova" ...  
## $ Comp\_SM : chr "Leaseflow" "Rentopia" "Leaseflow" "Leaseflow" ...  
## $ Dist\_SM : num 10.6 1.4 6.8 0.4 5.3 0.5 3.5 3.6 1.4 4 ...

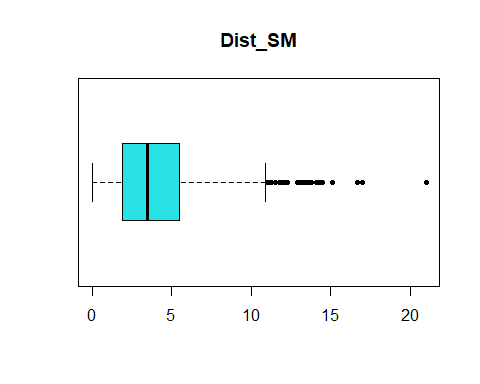
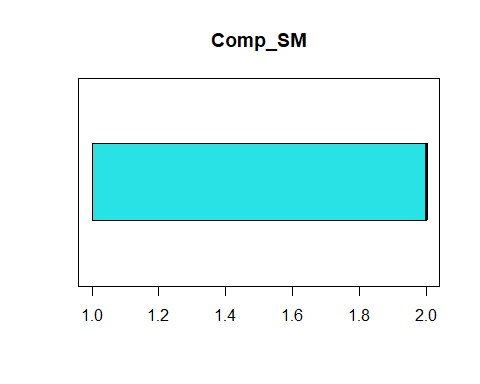
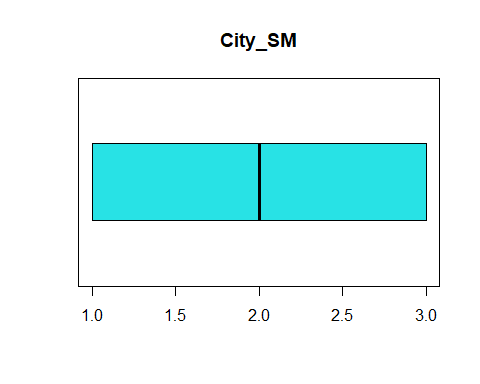
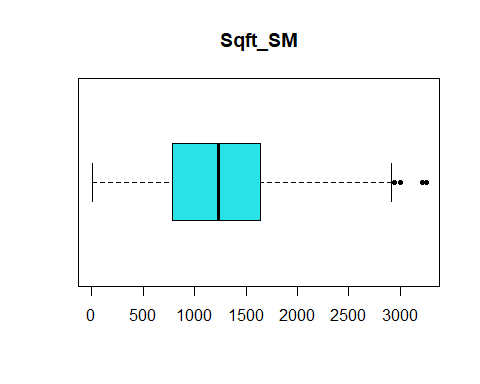
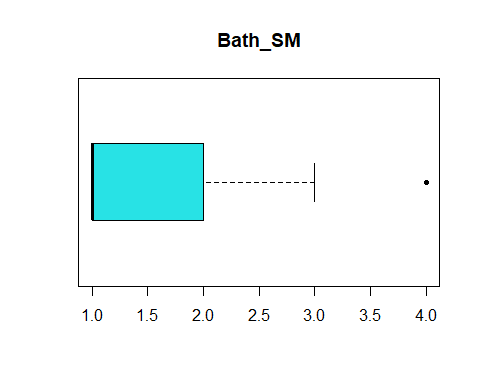
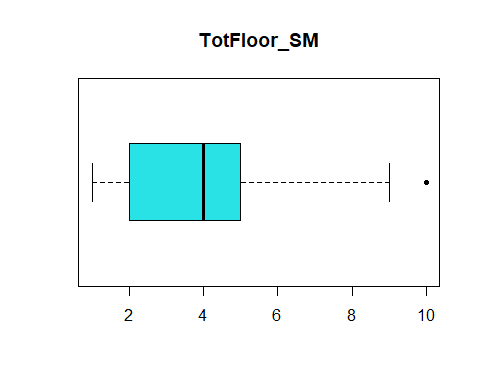
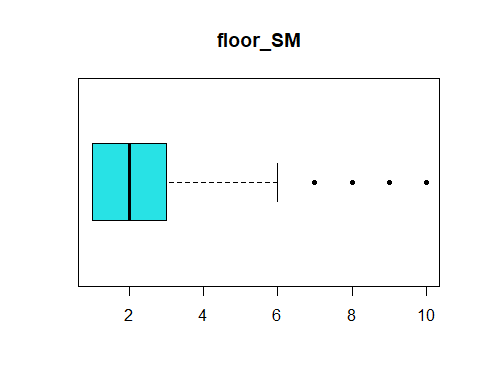
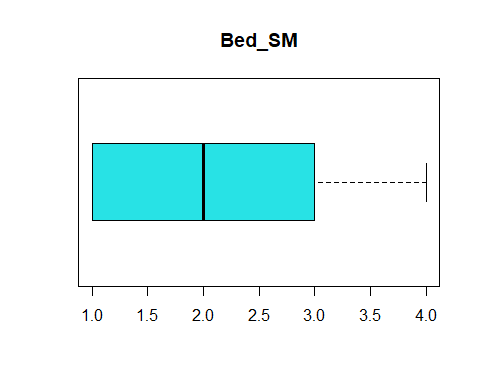
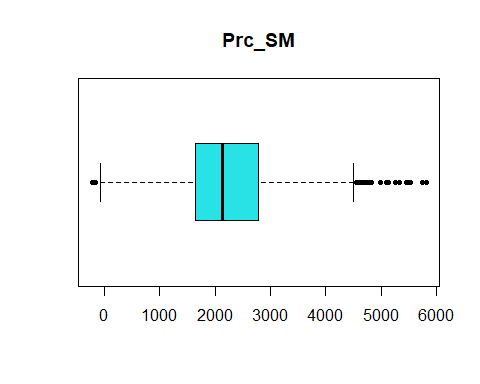
# Converting to factor   
df$City\_SM <- as.factor(df$City\_SM)  
df$Comp\_SM<- as.factor(df$Comp\_SM)  
  
  
kable(summary(df))

|  | Prc\_SM | Bed\_SM | floor\_SM | TotFloor\_SM | Bath\_SM | Sqft\_SM | City\_SM | Comp\_SM | Dist\_SM |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Min. :-218 | Min. :1.000 | Min. : 1.000 | Min. : 1.000 | Min. :1.000 | Min. : 1.0 | Blossomville:365 | Leaseflow:362 | Min. : 0.00 |
|  | 1st Qu.:1640 | 1st Qu.:1.000 | 1st Qu.: 1.000 | 1st Qu.: 2.000 | 1st Qu.:1.000 | 1st Qu.: 781.5 | Riverport :355 | Rentopia :689 | 1st Qu.: 1.90 |
|  | Median :2140 | Median :2.000 | Median : 2.000 | Median : 4.000 | Median :1.000 | Median :1231.0 | Terranova :331 | NA | Median : 3.50 |
|  | Mean :2260 | Mean :2.127 | Mean : 2.272 | Mean : 3.898 | Mean :1.618 | Mean :1289.1 | NA | NA | Mean : 4.11 |
|  | 3rd Qu.:2790 | 3rd Qu.:3.000 | 3rd Qu.: 3.000 | 3rd Qu.: 5.000 | 3rd Qu.:2.000 | 3rd Qu.:1645.0 | NA | NA | 3rd Qu.: 5.50 |
|  | Max. :5810 | Max. :4.000 | Max. :10.000 | Max. :10.000 | Max. :4.000 | Max. :3254.0 | NA | NA | Max. :21.00 |

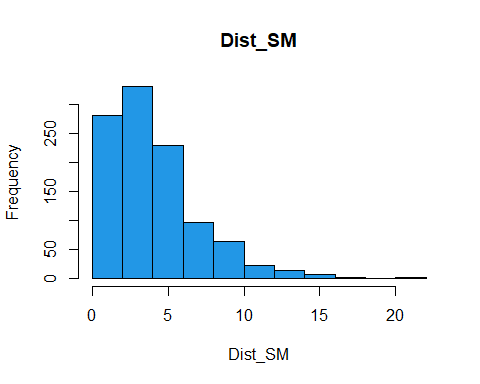
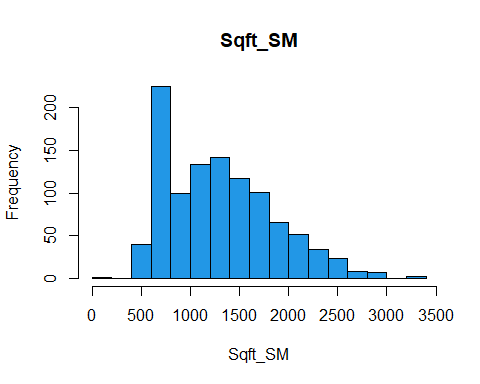
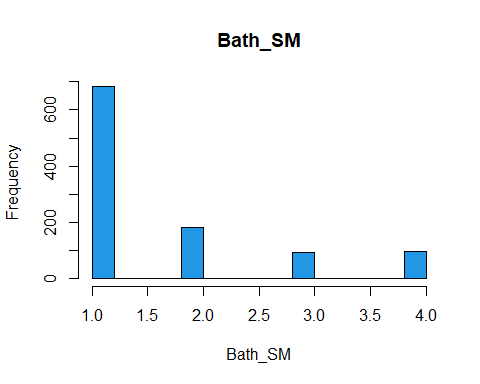
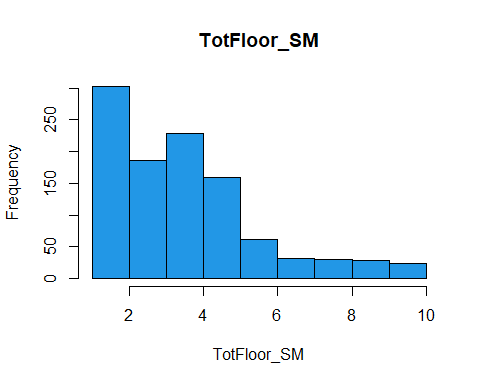
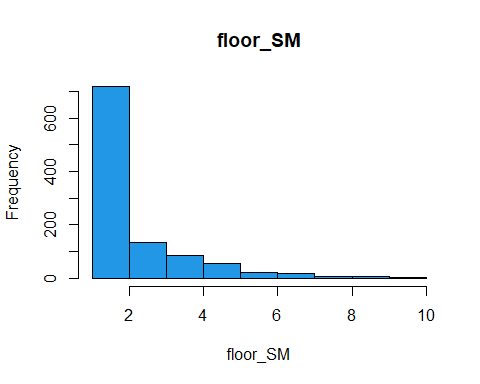
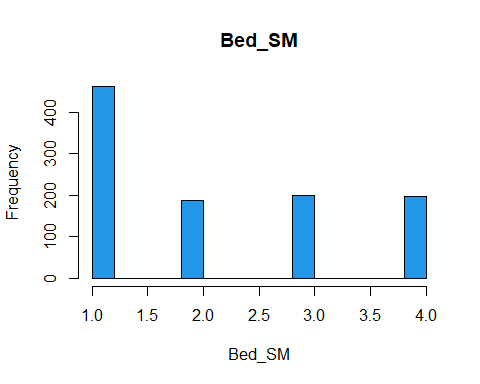
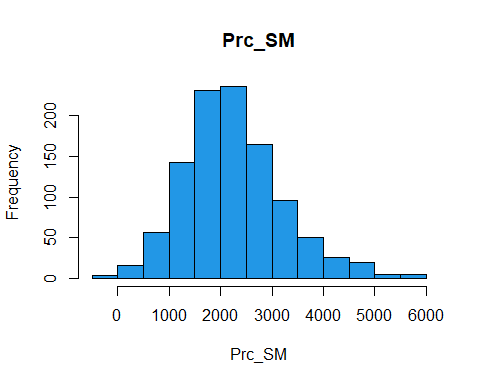
From the summary, there are some observations that are noteworthy. Prc\_SM has a min of -218, which could be an error as prices can not be negative Sqft\_SM (square feet) shows very low minimum value as well, indicates data error as well.

## Exploratory

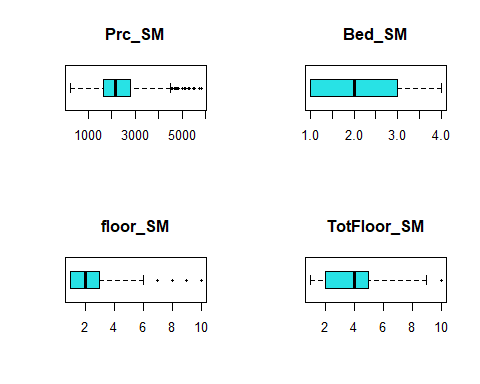
# Boxplots   
par(mfrow=c(1,1))  
for (i in 1:ncol(df)) {  
 boxplot(df[i], main=names(df)[i],  
 horizontal=TRUE, pch=20,col=5)  
}



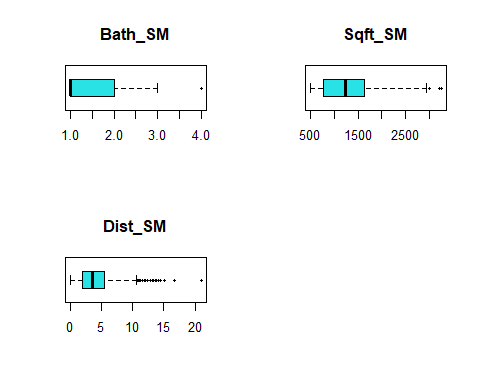
par(mfrow=c(1,1))  
  
#Histograms   
for (i in 1:ncol(df)) {  
 if (is.numeric(df[,i])) {  
 hist(df[,i], main=colnames(df)[i],  
 xlab=colnames(df)[i],col=4)  
 }  
}

 Boxplot of prices gave me a conviction and Histogram of Square feet made it clear of some data less than 500. Prices can not be negative as confirmed by the box plot and Square feet of an appartment can not be less than 500sqft, so we will be removing these values.

# Clearing the values   
df <- df[df$Prc\_SM > 0, ]  
  
df <- df[df$Sqft\_SM > 500, ]  
  
par(mfrow=c(2,2))  
for (i in 1:ncol(df)) {  
 if (is.numeric(df[,i])) {  
 boxplot(df[i], main=names(df)[i],  
 horizontal=TRUE, pch=20, col=5)  
 }  
}



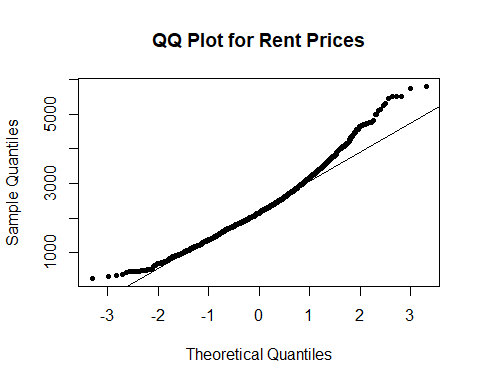
par(mfrow=c(1,1))

 ## 3. Rent Price Analysis between two companies

#Shapiro test  
shapiro.test(df$Prc\_SM)

##   
## Shapiro-Wilk normality test  
##   
## data: df$Prc\_SM  
## W = 0.97089, p-value = 0.0000000000001225

#QQ Normal   
qqnorm(df$Prc\_SM, main="QQ Plot for Rent Prices", pch=20)  
qqline(df$Prc\_SM)



# Variance F-Test  
var.test(Prc\_SM ~ Comp\_SM, data=df)

##   
## F test to compare two variances  
##   
## data: Prc\_SM by Comp\_SM  
## F = 1.1664, num df = 360, denom df = 684, p-value = 0.09053  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.9760439 1.4009961  
## sample estimates:  
## ratio of variances   
## 1.166375

From the shapiro test, p -value is less than 0.05 and the QQ plot does not show normality either, we can reject the null hypthosis meaning that the data do not follow a normal distribution and hence can not use the T-test.

We will go ahead with the Wilcox Test.

#Wilcoxon test  
wilcox.test(Prc\_SM ~ Comp\_SM, data=df)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: Prc\_SM by Comp\_SM  
## W = 120834, p-value = 0.5454  
## alternative hypothesis: true location shift is not equal to 0

Conclusion: Since the p-value > 0.05 we fail to reject the null hypothesis and can conclude that there is no evidence that shows rent prices differing from each other.

## 4. Training and Testing

# Find the number of rows of data  
n.row <- nrow(df)  
  
# Choose sampling rate  
sr <- 0.65  
  
# Last 4 digits of Student Number   
set.seed(4405)  
  
#Choose the rows for the training sample  
training.rows <- sample(1:n.row, sr\*n.row, replace=FALSE)  
  
#Assign to the training sample  
train <- subset(df[training.rows,])  
  
# Assign the balance to the Test Sample  
test <- subset(df[-c(training.rows),])

#Summary  
summary(train)

## Prc\_SM Bed\_SM floor\_SM TotFloor\_SM Bath\_SM   
## Min. : 255 Min. :1.000 Min. : 1.000 Min. : 1.00 Min. :1.00   
## 1st Qu.:1670 1st Qu.:1.000 1st Qu.: 1.000 1st Qu.: 2.00 1st Qu.:1.00   
## Median :2180 Median :2.000 Median : 2.000 Median : 4.00 Median :1.00   
## Mean :2280 Mean :2.084 Mean : 2.303 Mean : 3.91 Mean :1.61   
## 3rd Qu.:2785 3rd Qu.:3.000 3rd Qu.: 3.000 3rd Qu.: 5.00 3rd Qu.:2.00   
## Max. :5810 Max. :4.000 Max. :10.000 Max. :10.00 Max. :4.00   
## Sqft\_SM City\_SM Comp\_SM Dist\_SM   
## Min. : 501.0 Blossomville:238 Leaseflow:240 Min. : 0.000   
## 1st Qu.: 780.5 Riverport :229 Rentopia :439 1st Qu.: 1.900   
## Median :1223.0 Terranova :212 Median : 3.400   
## Mean :1277.7 Mean : 3.977   
## 3rd Qu.:1624.5 3rd Qu.: 5.200   
## Max. :3254.0 Max. :21.000

summary(test)

## Prc\_SM Bed\_SM floor\_SM TotFloor\_SM Bath\_SM   
## Min. : 373 Min. :1.000 Min. : 1.00 Min. : 1.000 Min. :1.00   
## 1st Qu.:1580 1st Qu.:1.000 1st Qu.: 1.00 1st Qu.: 2.000 1st Qu.:1.00   
## Median :2100 Median :2.000 Median : 2.00 Median : 4.000 Median :1.00   
## Mean :2247 Mean :2.196 Mean : 2.21 Mean : 3.872 Mean :1.64   
## 3rd Qu.:2810 3rd Qu.:3.000 3rd Qu.: 3.00 3rd Qu.: 5.000 3rd Qu.:2.00   
## Max. :5530 Max. :4.000 Max. :10.00 Max. :10.000 Max. :4.00   
## Sqft\_SM City\_SM Comp\_SM Dist\_SM   
## Min. : 501 Blossomville:125 Leaseflow:121 Min. : 0.200   
## 1st Qu.: 813 Riverport :124 Rentopia :246 1st Qu.: 2.100   
## Median :1241 Terranova :118 Median : 3.600   
## Mean :1317 Mean : 4.264   
## 3rd Qu.:1718 3rd Qu.: 5.800   
## Max. :2916 Max. :15.100

# Comparing means of each set   
round(mean(train$Prc\_SM),6)

## [1] 2280.184

round(mean(test$Prc\_SM),6)

## [1] 2246.575

# Using Wilcox for comparisson   
wilcox.test(train$Prc\_SM, test$Prc\_SM)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: train$Prc\_SM and test$Prc\_SM  
## W = 128132, p-value = 0.4484  
## alternative hypothesis: true location shift is not equal to 0

Conclusion: From the summary I could not find any evidence of dissimilarities which was further confirmed from Wilcox Test which showed p value= 0.4484 rejecting the null hypothesis.

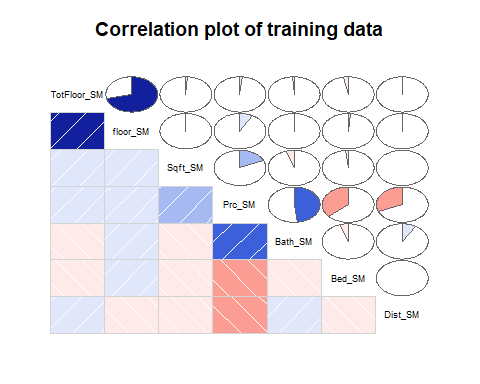
## 2. Simple Linear Regression

### 1. Correlations

# Numerical correlation  
train\_cor <- cor(train[sapply(train, is.numeric)], method="spearman")  
round(train\_cor, 2)

## Prc\_SM Bed\_SM floor\_SM TotFloor\_SM Bath\_SM Sqft\_SM Dist\_SM  
## Prc\_SM 1.00 -0.36 0.09 0.07 0.39 0.20 -0.29  
## Bed\_SM -0.36 1.00 0.01 -0.05 -0.04 0.00 -0.02  
## floor\_SM 0.09 0.01 1.00 0.67 0.01 0.01 -0.01  
## TotFloor\_SM 0.07 -0.05 0.67 1.00 -0.01 0.03 0.00  
## Bath\_SM 0.39 -0.04 0.01 -0.01 1.00 -0.03 0.09  
## Sqft\_SM 0.20 0.00 0.01 0.03 -0.03 1.00 0.00  
## Dist\_SM -0.29 -0.02 -0.01 0.00 0.09 0.00 1.00

# Graphical Correlation  
corrgram(train, order=TRUE, lower.panel=panel.shade,  
 upper.panel=panel.pie, text.panel=panel.txt,  
 main="Correlation plot of training data")

 Conclusions: floor\_sm (Floor) and totfloor\_sm (Total Floor) show a strong positive correlation which is obvious that higher floor number corresponds to appartment having more total floors.

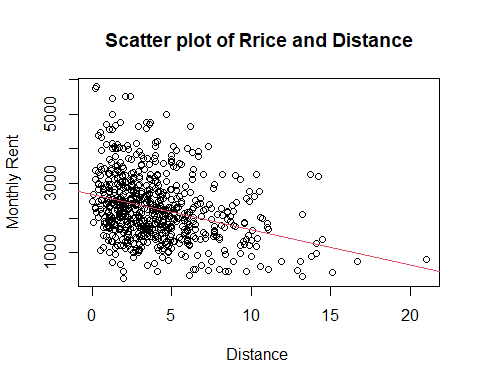
There are good observations in accordance to the prices Appartment Prices tend to increase and have moderate positive correlation (0.39) with Bathroom, which states more bathroom leads to higher prices

While having more Bedrooms leads to lower prices and has negative correlation (-0.36) between price and bedroom, which is quite common in real estates these days.

and it also decreases with the distance as seen between prc and dist (-0.29)

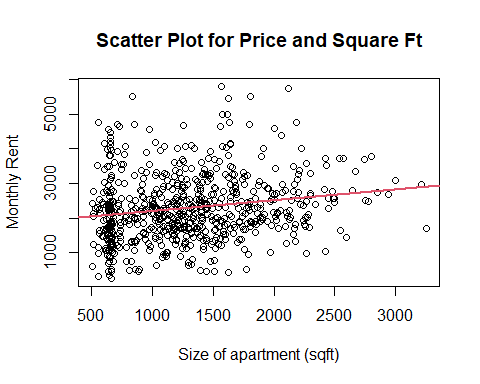
### 2. Simple linear regression model

# Creating linear model for Price and Distance   
mod.Dist <- lm(Prc\_SM ~ Dist\_SM, data = train)  
  
# Creating plot with regression line  
plot(Prc\_SM ~ Dist\_SM, data = train, pch = 1,  
 main = "Scatter plot of Rrice and Distance",  
 xlab = "Distance ",  
 ylab = "Monthly Rent")  
abline(mod.Dist, col = 2, lwd = 1)

 There is a an obvious negative relationship between the rent and the distance as shown the by the downward regression line. Shows high concentration of properites clustered withing the 0-5km distance with majority of rent prices from 1000 - 4000

### SLR for Price and Square Ft

# Creating the model  
mod.Sqft <- lm(Prc\_SM ~ Sqft\_SM, data = train)  
  
# Creating the plot with regression line  
plot(Prc\_SM ~ Sqft\_SM, data = train, pch = 1,  
 main = "Scatter Plot for Price and Square Ft",  
 xlab = "Size of apartment (sqft)",  
 ylab = "Monthly Rent")  
abline(mod.Sqft, col = 2, lwd = 2)

 Surpisingly there is not so significant rise in trend line in the indicating that the prices has minimal impact as the sqft of an appartment also increase, with a minimal positive correlation (0.29) as seen from the correlation matrix as well. The plots are quite varied throughout the chart with price variation occuring accross all sizes from 500 to 3000 square feet.

### 4. Comparing the two models

summary(mod.Dist)

##   
## Call:  
## lm(formula = Prc\_SM ~ Dist\_SM, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2226.2 -615.9 -101.2 445.6 3156.0   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2684.47 58.65 45.774 <2e-16 \*\*\*  
## Dist\_SM -101.65 11.96 -8.501 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 894.3 on 677 degrees of freedom  
## Multiple R-squared: 0.09646, Adjusted R-squared: 0.09512   
## F-statistic: 72.27 on 1 and 677 DF, p-value: < 2.2e-16

pred.Dist <- predict(mod.Dist, newdata=train)  
RMSE\_train\_Dist <- sqrt(mean((train$Prc\_SM - pred.Dist)^2))  
RMSE\_train\_Dist

## [1] 892.9838

summary(mod.Sqft)

##   
## Call:  
## lm(formula = Prc\_SM ~ Sqft\_SM, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1951.5 -605.3 -108.0 463.9 3439.6   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1881.79182 90.60897 20.768 < 2e-16 \*\*\*  
## Sqft\_SM 0.31180 0.06524 4.779 0.00000216 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 925.3 on 677 degrees of freedom  
## Multiple R-squared: 0.03264, Adjusted R-squared: 0.03121   
## F-statistic: 22.84 on 1 and 677 DF, p-value: 0.00000216

pred.Sqft <- predict(mod.Sqft, newdata=train)  
RMSE\_train\_Sqft <- sqrt(mean((train$Prc\_SM - pred.Sqft)^2))  
RMSE\_train\_Sqft

## [1] 923.9828

The first model is better than the second model as the Adjusted R-Squared is higher than that of the second model, which is out clear indicator.

# Comparing for Test Models   
pred.Dist\_test <- predict(mod.Dist, newdata=test)  
RMSE\_test\_Dist <- sqrt(mean((test$Prc\_SM - pred.Dist\_test)^2))  
RMSE\_test\_Dist

## [1] 911.8919

pred.Sqft\_tst <- predict(mod.Sqft, newdata=test)  
RMSE\_test\_Sqft <- sqrt(mean((test$Prc\_SM - pred.Sqft\_tst)^2))  
RMSE\_test\_Sqft

## [1] 940.6478

Model 1 has lower RMSE than the model 2 in the testing set so Model 1 is better.

Hence Model 1 is better in both Training and Testing set.

# 3. Model development - Multivariate

# Full Model  
full.model <- lm(Prc\_SM ~ . , data = train, na.action=na.omit)  
  
# Evaluating model on both the data  
summary(full.model)

##   
## Call:  
## lm(formula = Prc\_SM ~ ., data = train, na.action = na.omit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1926.40 -400.98 12.16 412.44 2115.40   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2036.48125 113.45028 17.950 < 2e-16 \*\*\*  
## Bed\_SM -291.97361 20.89330 -13.975 < 2e-16 \*\*\*  
## floor\_SM 76.62676 20.40129 3.756 0.000188 \*\*\*  
## TotFloor\_SM -36.89090 16.60434 -2.222 0.026633 \*   
## Bath\_SM 492.47161 25.19720 19.545 < 2e-16 \*\*\*  
## Sqft\_SM 0.32949 0.04421 7.453 0.000000000000283 \*\*\*  
## City\_SMRiverport 303.67631 58.34237 5.205 0.000000258299392 \*\*\*  
## City\_SMTerranova -194.36983 59.59658 -3.261 0.001165 \*\*   
## Comp\_SMRentopia 27.11305 50.49023 0.537 0.591449   
## Dist\_SM -113.91723 8.40264 -13.557 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 625.5 on 669 degrees of freedom  
## Multiple R-squared: 0.5632, Adjusted R-squared: 0.5574   
## F-statistic: 95.86 on 9 and 669 DF, p-value: < 2.2e-16

full.pred <- predict(full.model, newdata=train)  
RMSE\_train\_full <- sqrt(mean((train$Prc\_SM - full.pred)^2))  
RMSE\_train\_full

## [1] 620.8586

full.pred\_test <- predict(full.model, newdata = test)  
RMSE\_test\_full <- sqrt(mean((test$Prc\_SM - full.pred\_test)^2))  
RMSE\_test\_full

## [1] 614.3775

Comment on the main measures from the full model.

First the residuals shows median of 12.16 indicating some skewness in the model

P value of the coefficients is less than 0.05, with 8/9 variables aligning with intercept with p <0.05 and hence validating the prediction

We have Adjusted R-Square = 0.5574

F-stat with p-value (< 2.2e-16) which passes the hypothesis and model is significant

RMSE Train: 600.08 RMSE Test: 656.98

# Backward Selection Model  
back.model <- step(full.model, direction="backward", details=TRUE)

## Start: AIC=8753.44  
## Prc\_SM ~ Bed\_SM + floor\_SM + TotFloor\_SM + Bath\_SM + Sqft\_SM +   
## City\_SM + Comp\_SM + Dist\_SM  
##   
## Df Sum of Sq RSS AIC  
## - Comp\_SM 1 112816 261843839 8751.7  
## <none> 261731023 8753.4  
## - TotFloor\_SM 1 1931188 263662211 8756.4  
## - floor\_SM 1 5519182 267250206 8765.6  
## - Sqft\_SM 1 21732311 283463334 8805.6  
## - City\_SM 2 27353996 289085019 8816.9  
## - Dist\_SM 1 71907829 333638852 8916.3  
## - Bed\_SM 1 76401542 338132565 8925.3  
## - Bath\_SM 1 149446851 411177875 9058.1  
##   
## Step: AIC=8751.73  
## Prc\_SM ~ Bed\_SM + floor\_SM + TotFloor\_SM + Bath\_SM + Sqft\_SM +   
## City\_SM + Dist\_SM  
##   
## Df Sum of Sq RSS AIC  
## <none> 261843839 8751.7  
## - TotFloor\_SM 1 1919005 263762844 8754.7  
## - floor\_SM 1 5488832 267332672 8763.8  
## - Sqft\_SM 1 21721649 283565489 8803.8  
## - City\_SM 2 27718746 289562586 8816.1  
## - Dist\_SM 1 72041903 333885742 8914.8  
## - Bed\_SM 1 76289176 338133015 8923.3  
## - Bath\_SM 1 149409374 411253214 9056.3

# Evaluating the model  
summary(back.model)

##   
## Call:  
## lm(formula = Prc\_SM ~ Bed\_SM + floor\_SM + TotFloor\_SM + Bath\_SM +   
## Sqft\_SM + City\_SM + Dist\_SM, data = train, na.action = na.omit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1941.36 -396.71 13.25 407.84 2099.44   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2054.10016 108.54461 18.924 < 2e-16 \*\*\*  
## Bed\_SM -291.56969 20.86867 -13.972 < 2e-16 \*\*\*  
## floor\_SM 76.39931 20.38606 3.748 0.000194 \*\*\*  
## TotFloor\_SM -36.77103 16.59402 -2.216 0.027032 \*   
## Bath\_SM 491.79643 25.15245 19.553 < 2e-16 \*\*\*  
## Sqft\_SM 0.32940 0.04418 7.455 0.000000000000278 \*\*\*  
## City\_SMRiverport 305.66243 58.19409 5.252 0.000000201839318 \*\*\*  
## City\_SMTerranova -194.40966 59.56487 -3.264 0.001155 \*\*   
## Dist\_SM -114.00286 8.39667 -13.577 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 625.1 on 670 degrees of freedom  
## Multiple R-squared: 0.563, Adjusted R-squared: 0.5578   
## F-statistic: 107.9 on 8 and 670 DF, p-value: < 2.2e-16

back.pred <- predict(back.model, newdata=train)  
RMSE\_train\_back <- sqrt(mean((train$Prc\_SM - back.pred)^2))  
RMSE\_train\_back

## [1] 620.9924

back.pred\_tst <- predict(back.model, newdata = test)  
RMSE\_test\_back <- sqrt(mean((test$Prc\_SM - back.pred\_tst)^2))  
RMSE\_test\_back

## [1] 613.7087

Comment on the main measures from the backward model.

First the residuals shows median of 13.25 indicating some skewness in the model

P value of the coefficients is less than 0.05, with 8/8 that is all variables aligning with intercept with p <0.05 and hence validating the prediction

We have Adjusted R-Square = 0.5578

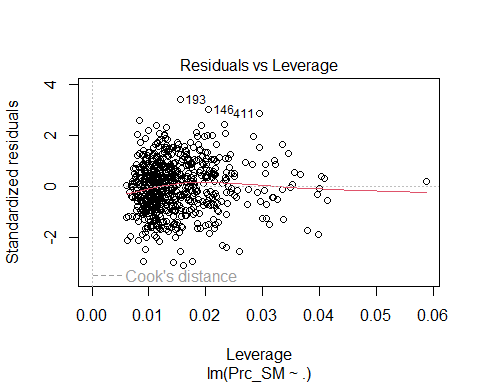
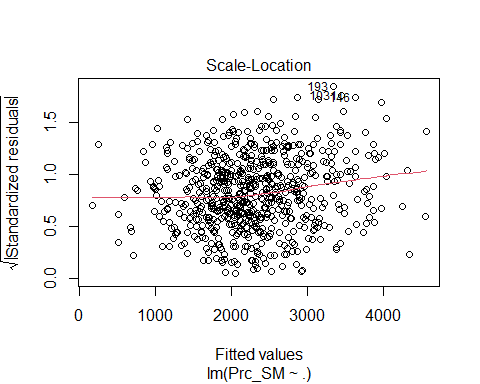
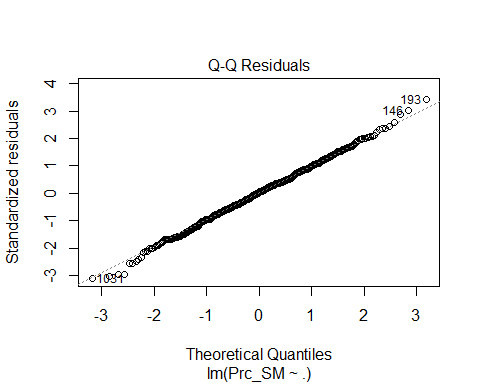
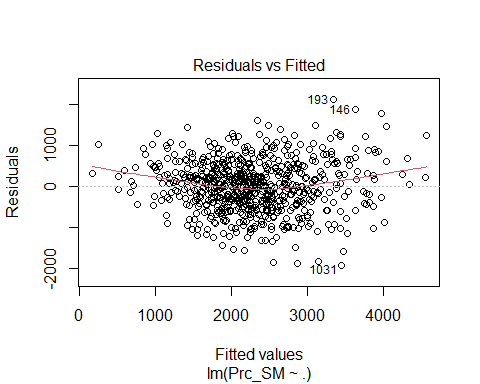
F-stat with p-value (< 2.2e-16) which passes the hypothesis and model is significant

RMSE Train: 620.9924 RMSE Test: 613.7087

Conclusion: Backwards Model is slightly better than the full model as Adjusted R-Squared is slightly higher (0.5578 vs 0.5574) with all vairables satisfying the intercept prediction.

# 4. Model evaluation

#Full Model   
plot(full.model)

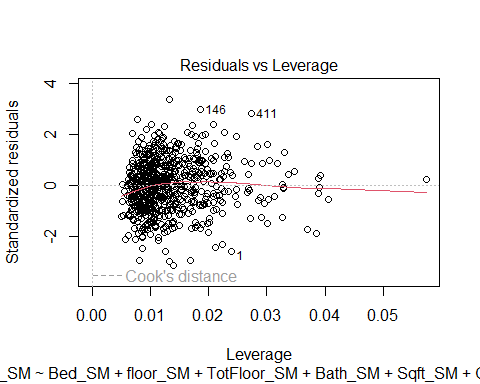
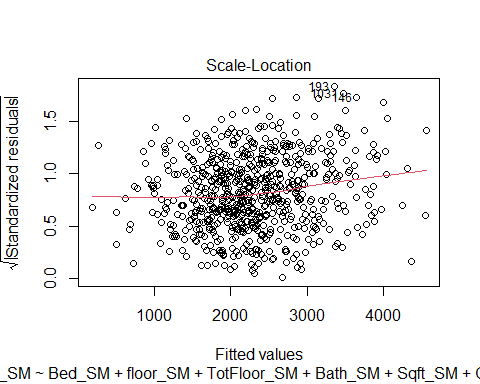
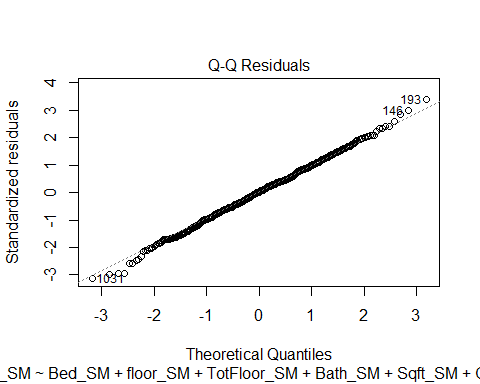
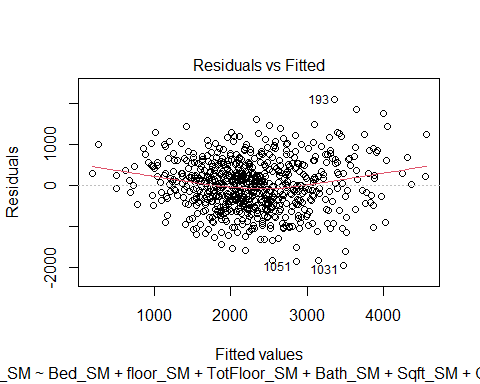
 Full Model:

Residuals vs Fitted - The points are scattered throughout the plot and no visible patterns. Satisfying the assumptions of linearity which can be proved further by looking at the scale-location plot.

QQ plot: Data points are following the line and only couple of points off from it so we can assume and suggest that the residuals are almost normally distributed.

Residuals vs Leverage Its crowded at on major spot with only few points going furhter, there are few points with high leverage but with no significant influence on it.

# Backwards Model   
plot(back.model)

 Evaluation of Backward Model:

The residuals vs fitted plot shows a relatively random scatter of points around zero.

QQ Normal PLot shows data points to be distributed on the line as well excpet for three data points which we can pass and assume that is also normally distributed.

There are not significant pattern observed the in the scatter plots so can conclude that it satisifies the assumptions of linearity

Residuals v Leverage plot : The trend line is relatively horizontal indicating a good lineraity, no points appear to exceed the Cook’s distance. There are couple of high leverage points (146 and 411) but without any influcence. Most of the data is clustered at particular point near 0.01 with only few points going outside.

Verifying the normaity

full.res <- residuals(full.model)  
back.res <- residuals(back.model)  
  
shapiro.test(full.res)

##   
## Shapiro-Wilk normality test  
##   
## data: full.res  
## W = 0.99853, p-value = 0.8516

shapiro.test(back.res)

##   
## Shapiro-Wilk normality test  
##   
## data: back.res  
## W = 0.99864, p-value = 0.8909

We can hence conclude that both models are normal.

#5. Final recommendation

After Analyzing both the models we can say that the Backwards Model that is the model 2 slightly better than the model 1 since it has lower RMSE value and slightly higher Adjusted R-squared with all variables satisfying the intercept prediction of p < 0.05

# References

1. PROG8435 – Data Analysis,Lecture 8,9 Prof. David Marsh All of my code was used from the sample code and lecture notes provided. ```