CS555\_term\_project\_Vu

Nguyen Vu

8/21/2019

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(stats)  
library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

library(tcltk)  
library(asbio)  
library(moments)  
library(ggplot2)  
library(ggcorrplot)

## Load Data File and Data Clean Up and Adding Variables

#### Location of data on Kaggle

#### <https://www.kaggle.com/goldenoakresearch/us-acs-mortgage-equity-loans-rent-statistics>

#load Data File  
  
#colums of interests:  
#c(1, 6, 7, 14:18, 22, 25, 37, 40, 47, 50, 52, 54)  
#read in giant CSV file (80 variables and 39030 observations):  
masterDT <- read.csv("real\_estate\_db.csv", header = T)  
# verify the names of the columns of interests above before subsetting:  
vars\_names <- names(masterDT[,c(1, 6, 7, 14:18, 22, 37, 47, 52)])  
vars\_names

## [1] "UID" "state" "state\_ab"   
## [4] "lat" "lng" "ALand"   
## [7] "AWater" "pop" "rent\_median"   
## [10] "hi\_median" "hc\_mortgage\_median" "hc\_median"

#subset columns that contains columns of interests and replace the original df with the smaller df  
masterDT <- masterDT[,c(1, 6, 7, 14:18, 22, 37, 47, 52)]#now data is reduced to 12 variables and 39030 observations  
#remove rows of incomplete cases  
masterDT <- masterDT[complete.cases(masterDT),] #data is reduced to 12 variables and 37942 observations  
#create a back up of the data frame incase there is an error during analysis.   
masterDT\_backup <- masterDT  
  
#rename columns to be more readable:  
new\_vars\_names <- gsub("hc", "Owner\_Costs", vars\_names)  
new\_vars\_names <- gsub("hi\_", "Household\_Income\_", new\_vars\_names)  
names(masterDT) <- new\_vars\_names  
  
#Create a varible to define the continental US vs states like Hawaii, Alaska and Puerto Rico to accurately  
#study at the effect longitude and latitude  
# Continental state = 1, non-continental states = 0  
masterDT$Continental <- ifelse(masterDT$state\_ab %in% c("HI", "AK", "PR"), 0, 1)  
  
# Create a value to define the the ratio of water to land ratio of specific geographic area  
masterDT$water\_land\_ratio <- masterDT$AWater/masterDT$ALand  
  
#creat a subset of just the continental US for analysis for Hawaii, Alaska and Puertorico can be considered geographically outliers  
contUSdf <- filter(masterDT, Continental !=0) #18 variables and 37228 observations  
attach(contUSdf)#attach to memory for ease of analysis  
  
# Let longitude of the US be devided into 3 equal segments.  
#Westcoast will be defined as lowest 1/3rd segment, Eastcoast will be the higest 1/3rd segment and Central will be the middle segment  
lng\_seg<- (max(contUSdf$lng)-min(contUSdf$lng))/3  
contUSdf$Coast <- cut(contUSdf$lng, breaks = c(-Inf, min(contUSdf$lng)+lng\_seg, min(contUSdf$lng)+2\*lng\_seg, Inf ),  
 labels = c("West", "Middle", "East"))  
  
#let latitude of the US be divided into 2 equal segments,  
# North will be defined as the highest half and the south will be defined as the lowest half.  
contUSdf$Region <- cut(contUSdf$lat, breaks = c(-Inf, mean(contUSdf$lat), Inf),   
 labels = c("South", "North"))  
  
  
# Calculate for the Population Density by divigin population by land Area  
contUSdf$Pop\_density <- contUSdf$pop/contUSdf$ALand

## Simple look at the data

Get the top 3 and bottom 3 states in term of highest and lowest gross rental median and owner’s total costs including mortgage

Get the average median rent, owner’s costs including mortgage

#Which State has the lowest average median cost for owner interm of Mortgage and cost?  
s<- contUSdf %>% group\_by(state) %>% summarise(avg\_mortgage\_plus\_cost = mean(Owner\_Costs\_mortgage\_median),   
 avg\_med\_rent = mean(rent\_median))  
s %>% arrange(avg\_mortgage\_plus\_cost)%>% slice(1:3)

## # A tibble: 3 x 3  
## state avg\_mortgage\_plus\_cost avg\_med\_rent  
## <fct> <dbl> <dbl>  
## 1 West Virginia 965. 642.  
## 2 Arkansas 1004. 677.  
## 3 Mississippi 1025. 714.

#which State has the highest average median cost for owner of Mortgage and cost?  
s %>% arrange(desc(avg\_mortgage\_plus\_cost))%>% slice(1:3)

## # A tibble: 3 x 3  
## state avg\_mortgage\_plus\_cost avg\_med\_rent  
## <fct> <dbl> <dbl>  
## 1 New Jersey 2378. 1351.  
## 2 New York 2177. 1224.  
## 3 District of Columbia 2176. 1366.

#Which state has the lowest average median gross median rent?  
s %>% arrange(avg\_med\_rent)%>% slice(1:3)

## # A tibble: 3 x 3  
## state avg\_mortgage\_plus\_cost avg\_med\_rent  
## <fct> <dbl> <dbl>  
## 1 South Dakota 1142. 625.  
## 2 West Virginia 965. 642.  
## 3 North Dakota 1138. 676.

#Which state has the highest average median gross median rent?  
s %>% arrange(desc(avg\_med\_rent))%>% slice(1:3)

## # A tibble: 3 x 3  
## state avg\_mortgage\_plus\_cost avg\_med\_rent  
## <fct> <dbl> <dbl>  
## 1 California 2165. 1421.  
## 2 Maryland 1922. 1370.  
## 3 District of Columbia 2176. 1366.

# get a table of averages summary of the average of Rent, Owner Cost, Owner Costs Mortgage and household Income by Coasts and region  
contUSdf%>% group\_by(Coast, Region) %>% summarise(US\_avg\_median\_rent = mean(rent\_median),  
 US\_avg\_median\_Owner\_Costs = mean(Owner\_Costs\_median),  
 US\_avg\_median\_Owner\_Costs\_n\_Mortgage = mean(Owner\_Costs\_mortgage\_median),  
 US\_avg\_median\_householdIncome = mean(Household\_Income\_median))

## # A tibble: 6 x 6  
## # Groups: Coast [3]  
## Coast Region US\_avg\_median\_r… US\_avg\_median\_O… US\_avg\_median\_O…  
## <fct> <fct> <dbl> <dbl> <dbl>  
## 1 West South 1317. 485. 1929.  
## 2 West North 1075. 483. 1699.  
## 3 Midd… South 845. 435. 1244.  
## 4 Midd… North 880. 510. 1408.  
## 5 East South 927. 429. 1327.  
## 6 East North 1061. 625. 1716.  
## # … with 1 more variable: US\_avg\_median\_householdIncome <dbl>

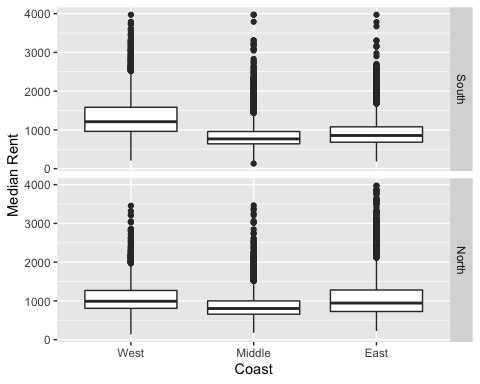
## More analysis by Location…

From the table above, it looks like the average median rent is the roughly the similar between regions (north vs south) but may be different between coasts (East, Middle and West). Check for this claim at the 95%CI

attach(contUSdf)

## The following objects are masked from contUSdf (pos = 3):  
##   
## ALand, AWater, Continental, Household\_Income\_median, lat, lng,  
## Owner\_Costs\_median, Owner\_Costs\_mortgage\_median, pop,  
## rent\_median, state, state\_ab, UID, water\_land\_ratio

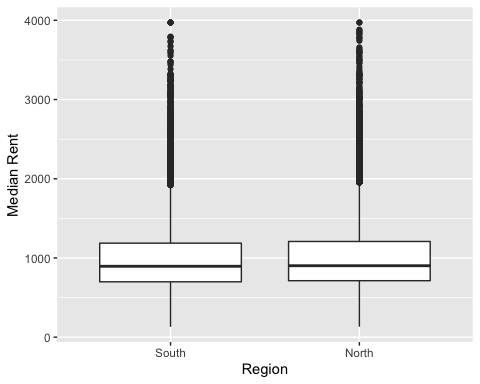
p <- ggplot(data = contUSdf) #load data in a plot  
# Graph rent distribution by boxplot  
p + geom\_boxplot(aes(x = Coast, y = rent\_median)) + facet\_grid(Region~.) + xlab("Coast") + ylab("Median Rent")



#### Gross Rent Analysis by Location: East VS West VS Middle coasts and North VS South region  
#Compare the average median rent the Northern VS. Sourthern US, and East Coast Vs. West Coast VS. Middle at the 95% CI:  
  
# Global F test looking at only Region:  
summary(aov(rent\_median~Region)) #not significant in global F test.

## Df Sum Sq Mean Sq F value Pr(>F)  
## Region 1 5.113e+05 511301 2.685 0.101  
## Residuals 37226 7.088e+09 190399

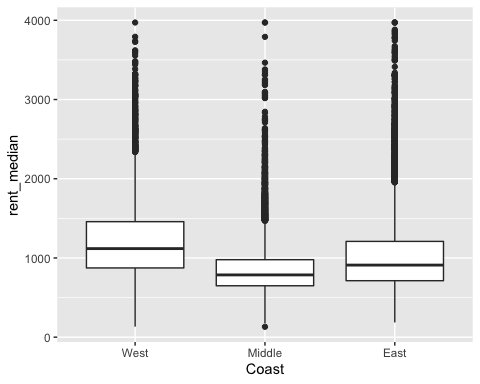
p + geom\_boxplot(aes(x = Region, y = rent\_median)) + ylab("Median Rent") #Region ONly Plot



#Compare the average median rent by coasts alone  
summary(aov(rent\_median~Coast)) #Global F Test is significant

## Df Sum Sq Mean Sq F value Pr(>F)   
## Coast 2 5.875e+08 293726116 1682 <2e-16 \*\*\*  
## Residuals 37225 6.501e+09 174637   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

p + geom\_boxplot(aes(x = Coast, y = rent\_median)) #Coast Only Plot



pairwise.t.test(rent\_median,Coast) #differences accross East Coasts, West Coast and Middle US.

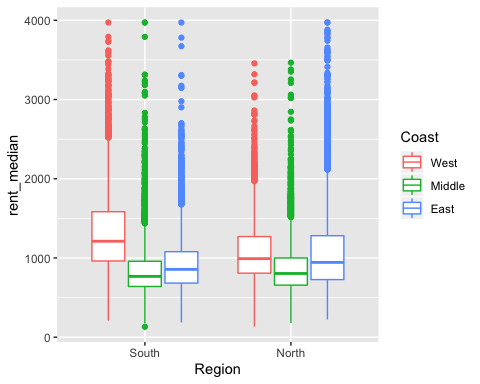
##   
## Pairwise comparisons using t tests with pooled SD   
##   
## data: rent\_median and Coast   
##   
## West Middle  
## Middle <2e-16 -   
## East <2e-16 <2e-16  
##   
## P value adjustment method: holm

No difference when comparing regions.  
F test indicate there is a difference between coasts and pairwise t test confirms that there is a significant diffence between the 3 coasts: East, West and Middle. Perform Ancova to check for covariance when fitting region and coasts. Check for interaction and fit model.

#Fit Rent when control for both Coasts and Region  
m\_test<-lm(rent\_median~Coast\*Region)  
Anova(m\_test, type = 3) #There is Interactions effect accross Groups

## Anova Table (Type III tests)  
##   
## Response: rent\_median  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 7896044777 1 46504.52 < 2.2e-16 \*\*\*  
## Coast 627976332 2 1849.26 < 2.2e-16 \*\*\*  
## Region 104012920 1 612.59 < 2.2e-16 \*\*\*  
## Coast:Region 175512772 2 516.85 < 2.2e-16 \*\*\*  
## Residuals 6319957312 37222   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#box plot of Coasts with Region by Color  
p + geom\_boxplot(aes(x = Region, y = rent\_median, colour = Coast))



#split data up by Region to stratify data to control for interaction

Interaction between coasts and regions exists. Stratifying is needed. Control for Region, test for difference between coast. Use West as the reference group.

m\_north <- lm(rent\_median[Region=="North"] ~ Coast[Region == "North"])  
m\_south <- lm(rent\_median[Region=="South"] ~ Coast[Region == "South"])  
Anova(m\_north, type = 3)#F significant

## Anova Table (Type III tests)  
##   
## Response: rent\_median[Region == "North"]  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 3396629607 1 19111.73 < 2.2e-16 \*\*\*  
## Coast[Region == "North"] 139873787 2 393.51 < 2.2e-16 \*\*\*  
## Residuals 3689389560 20759   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Anova(m\_south, type = 3)#F Significant

## Anova Table (Type III tests)  
##   
## Response: rent\_median[Region == "South"]  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 7896044777 1 49416 < 2.2e-16 \*\*\*  
## Coast[Region == "South"] 627976332 2 1965 < 2.2e-16 \*\*\*  
## Residuals 2630567753 16463   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#West is the reference group  
fitted\_m\_north <- fitted(m\_north)  
summary(m\_north) #in the north, no different in rent between West and East

##   
## Call:  
## lm(formula = rent\_median[Region == "North"] ~ Coast[Region ==   
## "North"])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -942.41 -292.16 -97.16 185.78 2910.84   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1075.405 7.779 138.245 <2e-16 \*\*\*  
## Coast[Region == "North"]Middle -194.930 9.572 -20.365 <2e-16 \*\*\*  
## Coast[Region == "North"]East -14.248 8.671 -1.643 0.1   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 421.6 on 20759 degrees of freedom  
## Multiple R-squared: 0.03653, Adjusted R-squared: 0.03643   
## F-statistic: 393.5 on 2 and 20759 DF, p-value: < 2.2e-16

# In the north, there is a difference between Middle and West   
  
fitted\_m\_south <- fitted(m\_south)  
summary(m\_south) # in the south, rent is difference between middle and east and west

##   
## Call:  
## lm(formula = rent\_median[Region == "South"] ~ Coast[Region ==   
## "South"])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1108.77 -247.96 -79.02 159.04 3126.98   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1316.765 5.923 222.30 <2e-16 \*\*\*  
## Coast[Region == "South"]Middle -471.744 7.919 -59.57 <2e-16 \*\*\*  
## Coast[Region == "South"]East -389.808 7.821 -49.84 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 399.7 on 16463 degrees of freedom  
## Multiple R-squared: 0.1927, Adjusted R-squared: 0.1926   
## F-statistic: 1965 on 2 and 16463 DF, p-value: < 2.2e-16

## Multivariate Regression Analysis of Median Gross Rent…

### Correlations

It would be intersting to look at the other variables available to see if we can estimate any other factor that influence Rent

attach(contUSdf)

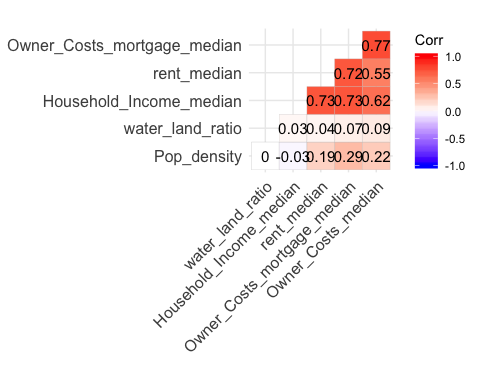
## The following objects are masked from contUSdf (pos = 3):  
##   
## ALand, AWater, Coast, Continental, Household\_Income\_median,  
## lat, lng, Owner\_Costs\_median, Owner\_Costs\_mortgage\_median,  
## pop, Pop\_density, Region, rent\_median, state, state\_ab, UID,  
## water\_land\_ratio

## The following objects are masked from contUSdf (pos = 4):  
##   
## ALand, AWater, Continental, Household\_Income\_median, lat, lng,  
## Owner\_Costs\_median, Owner\_Costs\_mortgage\_median, pop,  
## rent\_median, state, state\_ab, UID, water\_land\_ratio

# Graphicsal Correlations analysis of variables  
contUSdfFactors <- data.frame(Pop\_density, water\_land\_ratio, Household\_Income\_median, rent\_median, Owner\_Costs\_mortgage\_median, Owner\_Costs\_median)  
  
#Check for correlation  
cor(contUSdfFactors)

## Pop\_density water\_land\_ratio  
## Pop\_density 1.000000000 -0.001094935  
## water\_land\_ratio -0.001094935 1.000000000  
## Household\_Income\_median -0.033659606 0.030770360  
## rent\_median 0.188539513 0.042373239  
## Owner\_Costs\_mortgage\_median 0.291020159 0.070385853  
## Owner\_Costs\_median 0.219293568 0.088444516  
## Household\_Income\_median rent\_median  
## Pop\_density -0.03365961 0.18853951  
## water\_land\_ratio 0.03077036 0.04237324  
## Household\_Income\_median 1.00000000 0.72585094  
## rent\_median 0.72585094 1.00000000  
## Owner\_Costs\_mortgage\_median 0.72731369 0.72490791  
## Owner\_Costs\_median 0.61973600 0.55318778  
## Owner\_Costs\_mortgage\_median Owner\_Costs\_median  
## Pop\_density 0.29102016 0.21929357  
## water\_land\_ratio 0.07038585 0.08844452  
## Household\_Income\_median 0.72731369 0.61973600  
## rent\_median 0.72490791 0.55318778  
## Owner\_Costs\_mortgage\_median 1.00000000 0.77418206  
## Owner\_Costs\_median 0.77418206 1.00000000

c <- round(cor(contUSdfFactors), 2)  
ggcorrplot(c, type = "lower", lab = T)



Strong correlation:  
\* House hold income and Rent  
\* House hold income and Owner Cost + Mortgage  
\* House hold income and Owner Cost  
\* Owner Cost + Mortgage and Owner Cost

Moderate correlation for Owner Cost and Rent

Weak or no correlation between Population Density, and water:land ratio

### MlR analysis of median rent

# MLR for Rental median  
  
#Stratify to control for north south region because of interaction found in the previous section  
contUSdf\_N <- contUSdf[Region=="North",]  
contUSdf\_S <- contUSdf[Region=="South",]  
# Create dummy Variables for Coasts and Region for multiple regression analysis  
contUSdf\_N <- contUSdf %>% mutate(  
 gWest = if\_else(Coast=="West", 1, 0),  
 gMid = if\_else(Coast=="Middle", 1, 0),  
 gEast = if\_else(Coast=="East", 1, 0)  
)  
attach(contUSdf\_N)

## The following objects are masked from contUSdf (pos = 3):  
##   
## ALand, AWater, Coast, Continental, Household\_Income\_median,  
## lat, lng, Owner\_Costs\_median, Owner\_Costs\_mortgage\_median,  
## pop, Pop\_density, Region, rent\_median, state, state\_ab, UID,  
## water\_land\_ratio

## The following objects are masked from contUSdf (pos = 4):  
##   
## ALand, AWater, Coast, Continental, Household\_Income\_median,  
## lat, lng, Owner\_Costs\_median, Owner\_Costs\_mortgage\_median,  
## pop, Pop\_density, Region, rent\_median, state, state\_ab, UID,  
## water\_land\_ratio

## The following objects are masked from contUSdf (pos = 5):  
##   
## ALand, AWater, Continental, Household\_Income\_median, lat, lng,  
## Owner\_Costs\_median, Owner\_Costs\_mortgage\_median, pop,  
## rent\_median, state, state\_ab, UID, water\_land\_ratio

rent\_mlrm <- lm(rent\_median ~ Household\_Income\_median+  
 Pop\_density+water\_land\_ratio+  
 Owner\_Costs\_mortgage\_median+gWest+gMid)  
anova(rent\_mlrm) #F ratio is significant

## Analysis of Variance Table  
##   
## Response: rent\_median  
## Df Sum Sq Mean Sq F value  
## Household\_Income\_median 1 3734548651 3734548651 53982.407  
## Pop\_density 1 321868153 321868153 4652.562  
## water\_land\_ratio 1 2852555 2852555 41.233  
## Owner\_Costs\_mortgage\_median 1 318401620 318401620 4602.453  
## gWest 1 115504414 115504414 1669.601  
## gMid 1 20163025 20163025 291.454  
## Residuals 37221 2574980314 69181   
## Pr(>F)   
## Household\_Income\_median < 2.2e-16 \*\*\*  
## Pop\_density < 2.2e-16 \*\*\*  
## water\_land\_ratio 1.367e-10 \*\*\*  
## Owner\_Costs\_mortgage\_median < 2.2e-16 \*\*\*  
## gWest < 2.2e-16 \*\*\*  
## gMid < 2.2e-16 \*\*\*  
## Residuals   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(rent\_mlrm)

##   
## Call:  
## lm(formula = rent\_median ~ Household\_Income\_median + Pop\_density +   
## water\_land\_ratio + Owner\_Costs\_mortgage\_median + gWest +   
## gMid)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1886.80 -144.09 -15.24 127.28 2458.84   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.375e+02 3.985e+00 59.588 < 2e-16 \*\*\*  
## Household\_Income\_median 7.666e-03 7.500e-05 102.215 < 2e-16 \*\*\*  
## Pop\_density 1.116e+04 3.545e+02 31.478 < 2e-16 \*\*\*  
## water\_land\_ratio 9.173e+00 3.441e+00 2.666 0.00768 \*\*   
## Owner\_Costs\_mortgage\_median 1.903e-01 3.585e-03 53.075 < 2e-16 \*\*\*  
## gWest 1.253e+02 3.682e+00 34.041 < 2e-16 \*\*\*  
## gMid -5.464e+01 3.200e+00 -17.072 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 263 on 37221 degrees of freedom  
## Multiple R-squared: 0.6367, Adjusted R-squared: 0.6367   
## F-statistic: 1.087e+04 on 6 and 37221 DF, p-value: < 2.2e-16

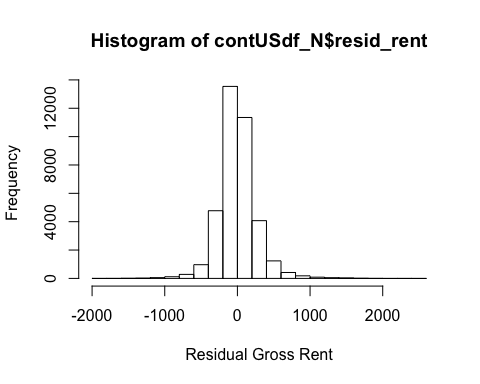
#R Square = 0.6367   
suma <- summary(rent\_mlrm)  
#get the slope estimates table for each predictor  
suma$coefficients

## Estimate Std. Error t value  
## (Intercept) 2.374837e+02 3.985421e+00 59.588112  
## Household\_Income\_median 7.665628e-03 7.499545e-05 102.214577  
## Pop\_density 1.115817e+04 3.544794e+02 31.477632  
## water\_land\_ratio 9.173288e+00 3.441095e+00 2.665804  
## Owner\_Costs\_mortgage\_median 1.902549e-01 3.584641e-03 53.075014  
## gWest 1.253356e+02 3.681850e+00 34.041461  
## gMid -5.463519e+01 3.200277e+00 -17.072020  
## Pr(>|t|)  
## (Intercept) 0.000000e+00  
## Household\_Income\_median 0.000000e+00  
## Pop\_density 1.160699e-214  
## water\_land\_ratio 7.683751e-03  
## Owner\_Costs\_mortgage\_median 0.000000e+00  
## gWest 3.794048e-250  
## gMid 4.245876e-65

#get 95% interval of slope estimates for each predictor  
confint(rent\_mlrm)

## 2.5 % 97.5 %  
## (Intercept) 2.296722e+02 2.452953e+02  
## Household\_Income\_median 7.518635e-03 7.812621e-03  
## Pop\_density 1.046338e+04 1.185296e+04  
## water\_land\_ratio 2.428645e+00 1.591793e+01  
## Owner\_Costs\_mortgage\_median 1.832289e-01 1.972809e-01  
## gWest 1.181190e+02 1.325521e+02  
## gMid -6.090783e+01 -4.836256e+01

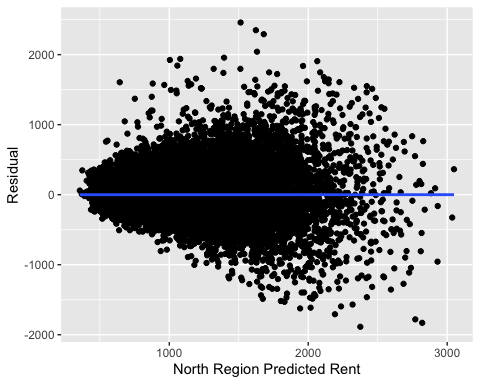
contUSdf\_N$fitted\_rent <- unclass(fitted(rent\_mlrm))  
contUSdf\_N$resid\_rent <- unclass(resid(rent\_mlrm))  
  
hist(contUSdf\_N$resid\_rent, xlab = "Residual Gross Rent")



skewness(contUSdf\_N$resid\_rent) #only slightly skewwed but resembled normal at skewness = 0.58

## [1] 0.6191456

ggplot(data = contUSdf\_N) + geom\_point(aes(x=fitted\_rent, y=resid\_rent))+ geom\_smooth(aes(x=fitted\_rent, y=resid\_rent), method = "lm", se = T) + xlab("North Region Predicted Rent")+ ylab("Residual")



#  
  
  
  
# Create dummy Variables for Coasts and Region for multiple regression analysis  
contUSdf\_S <- contUSdf %>% mutate(  
 gWest = if\_else(Coast=="West", 1, 0),  
 gMid = if\_else(Coast=="Middle", 1, 0),  
 gEast = if\_else(Coast=="East", 1, 0)  
)  
attach(contUSdf\_S)

## The following objects are masked from contUSdf\_N:  
##   
## ALand, AWater, Coast, Continental, gEast, gMid, gWest,  
## Household\_Income\_median, lat, lng, Owner\_Costs\_median,  
## Owner\_Costs\_mortgage\_median, pop, Pop\_density, Region,  
## rent\_median, state, state\_ab, UID, water\_land\_ratio

## The following objects are masked from contUSdf (pos = 4):  
##   
## ALand, AWater, Coast, Continental, Household\_Income\_median,  
## lat, lng, Owner\_Costs\_median, Owner\_Costs\_mortgage\_median,  
## pop, Pop\_density, Region, rent\_median, state, state\_ab, UID,  
## water\_land\_ratio

## The following objects are masked from contUSdf (pos = 5):  
##   
## ALand, AWater, Coast, Continental, Household\_Income\_median,  
## lat, lng, Owner\_Costs\_median, Owner\_Costs\_mortgage\_median,  
## pop, Pop\_density, Region, rent\_median, state, state\_ab, UID,  
## water\_land\_ratio

## The following objects are masked from contUSdf (pos = 6):  
##   
## ALand, AWater, Continental, Household\_Income\_median, lat, lng,  
## Owner\_Costs\_median, Owner\_Costs\_mortgage\_median, pop,  
## rent\_median, state, state\_ab, UID, water\_land\_ratio

rent\_mlrm2 <- lm(rent\_median ~ Household\_Income\_median+  
 Pop\_density+water\_land\_ratio+  
 Owner\_Costs\_mortgage\_median+gWest+gMid)  
  
summary(rent\_mlrm2)

##   
## Call:  
## lm(formula = rent\_median ~ Household\_Income\_median + Pop\_density +   
## water\_land\_ratio + Owner\_Costs\_mortgage\_median + gWest +   
## gMid)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1886.80 -144.09 -15.24 127.28 2458.84   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.375e+02 3.985e+00 59.588 < 2e-16 \*\*\*  
## Household\_Income\_median 7.666e-03 7.500e-05 102.215 < 2e-16 \*\*\*  
## Pop\_density 1.116e+04 3.545e+02 31.478 < 2e-16 \*\*\*  
## water\_land\_ratio 9.173e+00 3.441e+00 2.666 0.00768 \*\*   
## Owner\_Costs\_mortgage\_median 1.903e-01 3.585e-03 53.075 < 2e-16 \*\*\*  
## gWest 1.253e+02 3.682e+00 34.041 < 2e-16 \*\*\*  
## gMid -5.464e+01 3.200e+00 -17.072 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 263 on 37221 degrees of freedom  
## Multiple R-squared: 0.6367, Adjusted R-squared: 0.6367   
## F-statistic: 1.087e+04 on 6 and 37221 DF, p-value: < 2.2e-16

anova(rent\_mlrm2)

## Analysis of Variance Table  
##   
## Response: rent\_median  
## Df Sum Sq Mean Sq F value  
## Household\_Income\_median 1 3734548651 3734548651 53982.407  
## Pop\_density 1 321868153 321868153 4652.562  
## water\_land\_ratio 1 2852555 2852555 41.233  
## Owner\_Costs\_mortgage\_median 1 318401620 318401620 4602.453  
## gWest 1 115504414 115504414 1669.601  
## gMid 1 20163025 20163025 291.454  
## Residuals 37221 2574980314 69181   
## Pr(>F)   
## Household\_Income\_median < 2.2e-16 \*\*\*  
## Pop\_density < 2.2e-16 \*\*\*  
## water\_land\_ratio 1.367e-10 \*\*\*  
## Owner\_Costs\_mortgage\_median < 2.2e-16 \*\*\*  
## gWest < 2.2e-16 \*\*\*  
## gMid < 2.2e-16 \*\*\*  
## Residuals   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

suma2 <- summary(rent\_mlrm2)  
#get the slope estimates table for each predictor  
suma2$coefficients

## Estimate Std. Error t value  
## (Intercept) 2.374837e+02 3.985421e+00 59.588112  
## Household\_Income\_median 7.665628e-03 7.499545e-05 102.214577  
## Pop\_density 1.115817e+04 3.544794e+02 31.477632  
## water\_land\_ratio 9.173288e+00 3.441095e+00 2.665804  
## Owner\_Costs\_mortgage\_median 1.902549e-01 3.584641e-03 53.075014  
## gWest 1.253356e+02 3.681850e+00 34.041461  
## gMid -5.463519e+01 3.200277e+00 -17.072020  
## Pr(>|t|)  
## (Intercept) 0.000000e+00  
## Household\_Income\_median 0.000000e+00  
## Pop\_density 1.160699e-214  
## water\_land\_ratio 7.683751e-03  
## Owner\_Costs\_mortgage\_median 0.000000e+00  
## gWest 3.794048e-250  
## gMid 4.245876e-65

#get 95% interval of slope estimates for each predictor  
confint(rent\_mlrm2)

## 2.5 % 97.5 %  
## (Intercept) 2.296722e+02 2.452953e+02  
## Household\_Income\_median 7.518635e-03 7.812621e-03  
## Pop\_density 1.046338e+04 1.185296e+04  
## water\_land\_ratio 2.428645e+00 1.591793e+01  
## Owner\_Costs\_mortgage\_median 1.832289e-01 1.972809e-01  
## gWest 1.181190e+02 1.325521e+02  
## gMid -6.090783e+01 -4.836256e+01

#get predicted and residuals  
contUSdf\_S$fitted\_rent <- unclass(fitted(rent\_mlrm2))  
contUSdf\_S$resid\_rent <- unclass(resid(rent\_mlrm2))  
attach(contUSdf\_S)

## The following objects are masked from contUSdf\_S (pos = 3):  
##   
## ALand, AWater, Coast, Continental, gEast, gMid, gWest,  
## Household\_Income\_median, lat, lng, Owner\_Costs\_median,  
## Owner\_Costs\_mortgage\_median, pop, Pop\_density, Region,  
## rent\_median, state, state\_ab, UID, water\_land\_ratio

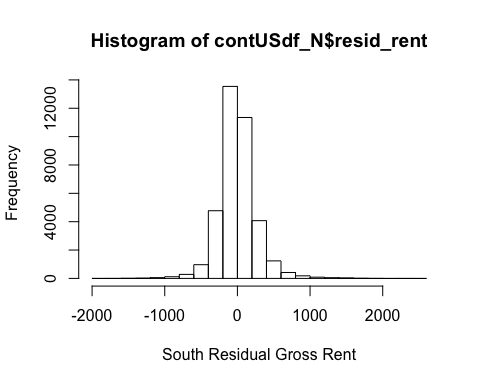
## The following objects are masked from contUSdf\_N:  
##   
## ALand, AWater, Coast, Continental, gEast, gMid, gWest,  
## Household\_Income\_median, lat, lng, Owner\_Costs\_median,  
## Owner\_Costs\_mortgage\_median, pop, Pop\_density, Region,  
## rent\_median, state, state\_ab, UID, water\_land\_ratio

## The following objects are masked from contUSdf (pos = 5):  
##   
## ALand, AWater, Coast, Continental, Household\_Income\_median,  
## lat, lng, Owner\_Costs\_median, Owner\_Costs\_mortgage\_median,  
## pop, Pop\_density, Region, rent\_median, state, state\_ab, UID,  
## water\_land\_ratio

## The following objects are masked from contUSdf (pos = 6):  
##   
## ALand, AWater, Coast, Continental, Household\_Income\_median,  
## lat, lng, Owner\_Costs\_median, Owner\_Costs\_mortgage\_median,  
## pop, Pop\_density, Region, rent\_median, state, state\_ab, UID,  
## water\_land\_ratio

## The following objects are masked from contUSdf (pos = 7):  
##   
## ALand, AWater, Continental, Household\_Income\_median, lat, lng,  
## Owner\_Costs\_median, Owner\_Costs\_mortgage\_median, pop,  
## rent\_median, state, state\_ab, UID, water\_land\_ratio

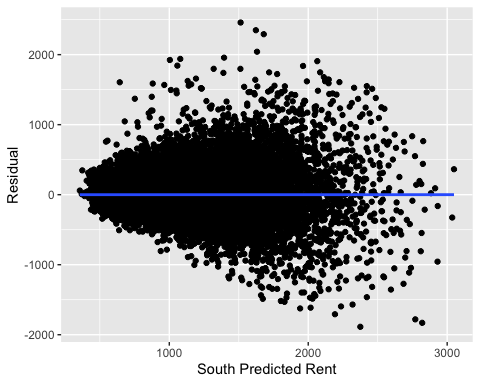
hist(contUSdf\_N$resid\_rent, xlab = "South Residual Gross Rent")



skewness(contUSdf\_N$resid\_rent) #only slightly skewwed but resembled normal at skewness = 0.62

## [1] 0.6191456

ggplot(data = contUSdf\_S) + geom\_point(aes(x=fitted\_rent, y=resid\_rent))+ geom\_smooth(aes(x=fitted\_rent, y=resid\_rent), method = "lm", se = T) + xlab("South Predicted Rent")+ ylab("Residual")



In both region, South and North, we’ve found evidence that rental prices is in fluenced by all tested factors. Our residuals are normally shaped but not evenly distributed. There may be missing factors. From the residuals plot, it looks like our prediction is more accurate at the lower rental value than the higher ones.

Although every predictor variables has a significant effect on rental price, from the slope Estimates, it looks like median income has the largest slope change per unit on rental price.

## Proportion test of pricy rental above 90 percentiles of median rent

##############################  
##############################  
# Comparing the states with the highest proportion of rent above 75 percentile price.   
#get the 90 percentile of all the median rent in the US  
US90PercentileRent <- quantile(rent\_median, 0.90)  
  
# Create variable Top10% of area with median above the 90 percentile of the national median  
# 1 is above, 0 is below  
contUSdf$top10percent <- if\_else(rent\_median>US90PercentileRent, 1, 0)  
  
#get the proportion of homes by coasts with proportion of median rent in the top 10% of the nation  
contUSdf %>% group\_by(Coast) %>% count(top10percent==1)

## # A tibble: 6 x 3  
## # Groups: Coast [3]  
## Coast `top10percent == 1` n  
## <fct> <lgl> <int>  
## 1 West FALSE 6026  
## 2 West TRUE 1465  
## 3 Middle FALSE 11036  
## 4 Middle TRUE 462  
## 5 East FALSE 16450  
## 6 East TRUE 1789

contUSdf %>% group\_by(Coast) %>% summarise(n = n())

## # A tibble: 3 x 2  
## Coast n  
## <fct> <int>  
## 1 West 7491  
## 2 Middle 11498  
## 3 East 18239

#Test the hypotheis that the East Coast has just as % rental market with median prices higher than   
# 90% of the median as the west coasts  
  
prop.test(c(1789,1465), c(18239,7491), conf.level = 0.95)

##   
## 2-sample test for equality of proportions with continuity  
## correction  
##   
## data: c(1789, 1465) out of c(18239, 7491)  
## X-squared = 455.88, df = 1, p-value < 2.2e-16  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## -0.10754101 -0.08742198  
## sample estimates:  
## prop 1 prop 2   
## 0.09808652 0.19556801

#The % of rental in the East Coast to be outrageous pricing above the 90% of national median is actually less than the west by  
# 8.7% to 10%.

The % of rental in the East Coast to be outrageous pricing above the 90% of national median is actually less than the west by # 8.7% to 10%.