R Notebook

#packages

Housing Data Summary

House Pricing is the most intrinsic factor of economy, and they are of great interest for buyers and sellers. Moreover, nowadays housing price and property taxes is increasing rapidly and it is an important factor that needs to be considered, before purchasing a house because it is a long - term investment. From the survey of 2007 and 2008, it was found that many people bought the house on the basis of assumptions that housing price and property price will decrease in year 2007 and considering that factor, they took the loan from the bank and invested into properties, but that was not the case and this recession impacted many financial statements of individuals. Thus, the goal of the project is to build a regression model that would help to determine the factors that would lead to increase in housing price in different divisions and states.My aim is to focus on model which predict the Property Taxes on the basis of divisions, states, insurance, bathroom, kitchen and many more factors. This will give consciousness to individuals about the considerations of factors that needs to be taken before buying or selling house.

Methodology

Moreover,from ACS housing data I have filtered some columns which includes State,Division, Acres,Tax, Agro Products sales, Bath, Kitchen, Rent of house,and other household utilites which will help me to find the prediction of house before buying. I had similarly done this project in Foundation of Modelling in which we need to analyze some reserach paper on the basis of some topic. So, I decided to go with the Housing data in R and apply some research methods on it. In the research paper,they had simply cleaned the data and applied linear regression model on it, but in my project I tried to do some tests on it and also performed different reltion, which can help individual in buying housing property.

Therefore, I have performed different steps to identify regression model. 1st step: Loading and filtering data 2ndstep: Weighted mean of Monthly Rent and Insurance 3rd step: Labelling factors for better understanding 4th step: Performing different graphs for understanding relationship 5th step: Applied some tests on it 6th step: Checking AIC and Bic, to see which model fits better 7th step: Splitting data into train and test data 8th step: Building Linear Regression Model 9th step: Predicting model by using test data 10th step: Visualizing summary of model

```
fields <-  c("RT", "DIVISION","ADJHSG", "ST","WGTP",
             "ACR","AGS","BATH","BDSP", "HOTWAT",
             "INSP","RMSP", "SINK","STOV","TEL",
             "TOIL","VALP","YBL", "KIT","TAXP",
             "RNTP")


A <- data.frame(fread
("C:/Users/Janvi/Documents/R/Final Project/csv_hus/psam_husa.csv",
                      header=TRUE,select = fields))

B <- data.frame(fread
("C:/Users/Janvi/Documents/R/Final Project/csv_hus/psam_husb.csv",
                      header=TRUE,select = fields))

C <- data.frame(fread
("C:/Users/Janvi/Documents/R/Final Project/csv_hus/psam_husc.csv",
                      header=TRUE,select = fields))

D <- data.frame(fread
("C:/Users/Janvi/Documents/R/Final Project/csv_hus/psam_husd.csv",
                      header=TRUE,select = fields))

bind_data <- rbind(A,B,C,D)

bind_data <- bind_data %>%
  rename("RecordType" = RT,"DIVISION"= DIVISION,
         "Adjacent Factor" = ADJHSG,"State" = ST,
         "Housingweight" = WGTP,"HouseAcre" = ACR,
         "SaleofAgroProduct"= AGS,"Bathtub" = BATH,
         "Bedrooms" = BDSP,"HotWater" = HOTWAT,
         "Insurance" = INSP,"Stove" = STOV,
         "TelephoneService" = TEL,"Toilet" = TOIL,
         "PropertyValue" = VALP,"HouseStructureYear" = YBL,
         "Kitchen" = KIT,"Tax" = TAXP,"MonthlyRent" = RNTP)
View(bind_data)
```

Weighted mean and labelling factors

In this section I have weighted monthy rent by adjacent factor to result it into dollars, then I have done same for the Insurance. Furthermore, I have labelled the factors of state, year built in and divisions.In the end of this chunk I have omitted the Na values and based upon that I have performed different relation of graphs.

```
#Weighted monthly rent
bind_data["RENT"]=bind_data["Adjacent Factor"]*bind_data["MonthlyRent"]/1000000

###Weighted mean of Insurance
bind_data["INSURANCE"]=bind_data["Adjacent Factor"]*bind_data["Insurance"]/1000000

#Labeling factors of DIVISION
bind_data$DIVISION <- factor(bind_data$DIVISION,
                             levels = c(1,2,3,4,5,6,7,8,9),
                             labels = c("New England", "Middle Atlantic",
                                        "East North Central",
                                        "West North Central",
                                        "South Atlantic",
                                        "East South Central",
                                        "West South Central",
                                        "Mountain","Pacific"))
bind_data$HouseStructureYear <- factor(bind_data$HouseStructureYear,
                             levels = c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,
                                        16,17,18,19,20,21),
                             labels = c("1939 or earlier","1940 to 1949",
                                        "1950 to 1959",
                                        "1960 to 1969",
                                        "1970 to 1979",
                                        "1980 to 1989",
                                        "1990 to 1999",
                                        "2000 to 2004",
                                        "2005","2006","2007",
                                        "2008","2009","2010",
                                        "2011","2012","2013",
                                        "2014",
                                        "2015",
                                        "2016 ",
                                        "2017"))
#Labeling states
bind_data$State <- factor(bind_data$State,
                       levels = c(1,2,4,5,6,8,9,
                                  10,11,12,13,15,16,17,18,
                                  19,20,21,22,23,24,25,26,27,
                                  28,29,30,31,32,33,34,35,36,
                                  37,38,39,40,41,42,44,45,
                                  46,47,48,49,50,51,53,54,
                                  55,56,72),

                       labels =
                         c("AL","AK","AZ","AR","CA","CO","CT","DE",
                           "DC","FL","GA","HI","ID","IL","IN","IA",
                           "KS","KY","LA","ME","MD","MA","MI","MN",
                           "MS","MO","MT","NE","NV","NH","NJ","NM",
```

```
                             "NY","NC","ND","OH","OK","OR","PA","RI",
                             "SC","SD","TN","TX","UT","VT","VA","WA",
                             "WV","WI","WY","PR"))
##Removing NA from data
Without_NA <- bind_data %>% select(State,DIVISION,Bathtub,HotWater,Bedrooms,RMSP,SINK,
Stove,Toilet,
       HouseStructureYear,Kitchen,RENT) %>% group_by(RENT) %>% na.omit()
head(Without_NA)
```

State <fctr>	DIVISION <fctr>	Bathtub <int>	HotWater <int>	Bedro… <int>	R… <int>	S… <int>	St… <int>	Toilet <int>	House <fctr>
AL	East South Central	1	9	4	6	1	1	1	1980 to
AL	East South Central	1	9	1	3	1	1	1	1970 to
AL	East South Central	1	9	1	2	1	1	1	2006
AL	East South Central	1	9	2	4	1	1	1	1990 to
AL	East South Central	1	9	2	4	1	1	1	1980 to
AL	East South Central	1	9	3	6	1	1	1	1950 to

6 rows | 1-10 of 12 columns

Plotting Divisions by 2017 year wise

From the below graph we can see that number of houses built in South Atlantic are around 400 in year 2017 and least were built in New England, so the consumption of lands in New England is less, so we can predict that, rent in that division would be less. Moreover, when we tried that relation with omitted NA values then west south central shows highest built houses in year 2017, which is wrong prediction and thus by omitting NA can change a lot of result.

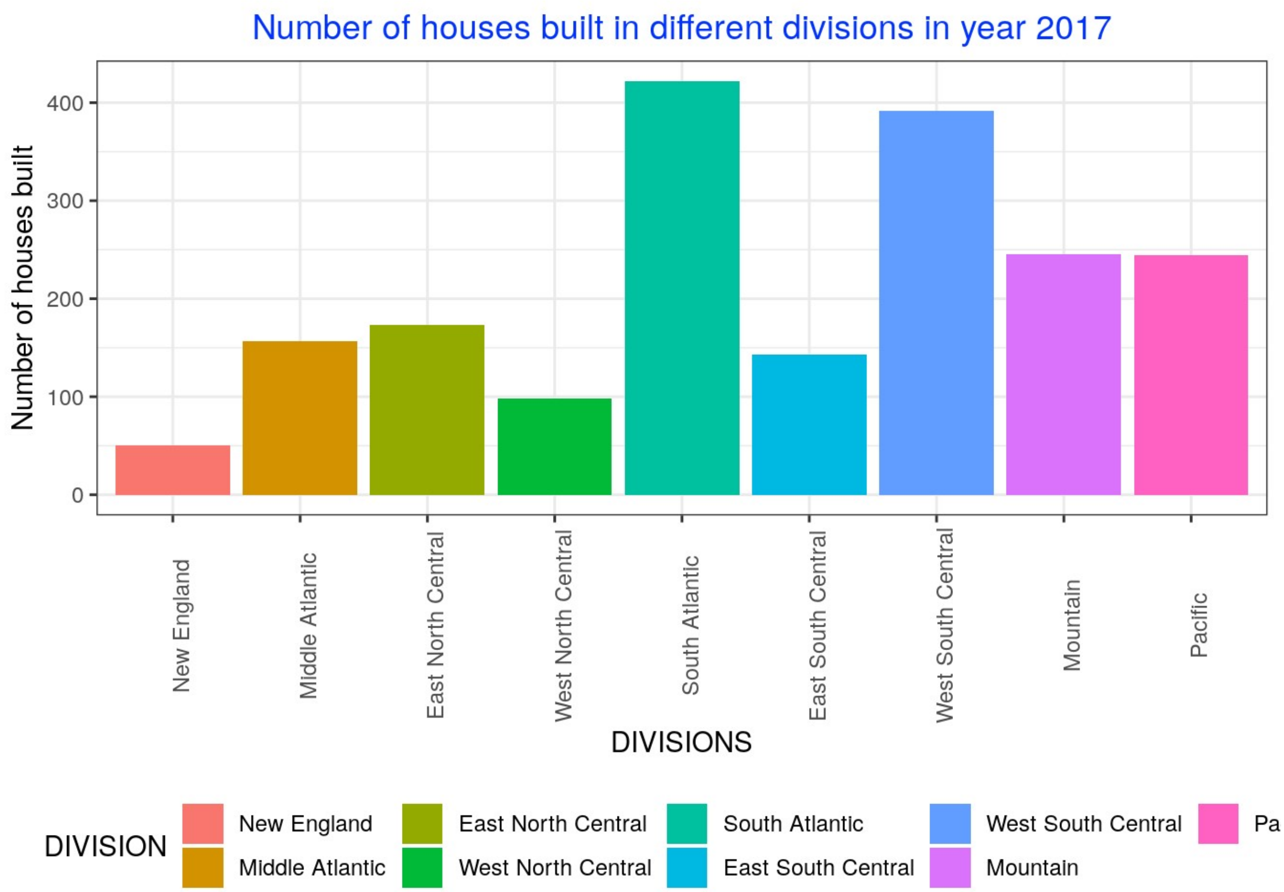
```
#Analysing Divisions Rent in 2017 year
year_division_bind <- bind_data %>%
  select(DIVISION,Bathtub,RMSP,HouseStructureYear,RENT) %>%
  filter(HouseStructureYear == 2017) %>%
  group_by(DIVISION) %>%
  summarise(Count=n())
year_division_bind
```

DIVISION <fctr>	Count <int>
New England	50
Middle Atlantic	157

DIVISION <fctr>	Count <int>
East North Central	173
West North Central	98
South Atlantic	422
East South Central	143
West South Central	392
Mountain	245
Pacific	244

9 rows

```
#plot
ggplot(year_division_bind)+
  geom_col(mapping =aes(x= DIVISION,y= Count,fill=DIVISION))+
  ggtitle("Number of houses built in different divisions in year 2017")+
  xlab("DIVISIONS")+ylab("Number of houses built")+ theme_bw()+
  theme(plot.title= element_text(color="#0033FF",hjust = 0.5),
        axis.text.x = element_text(angle = 90),
        legend.position= "bottom")
```



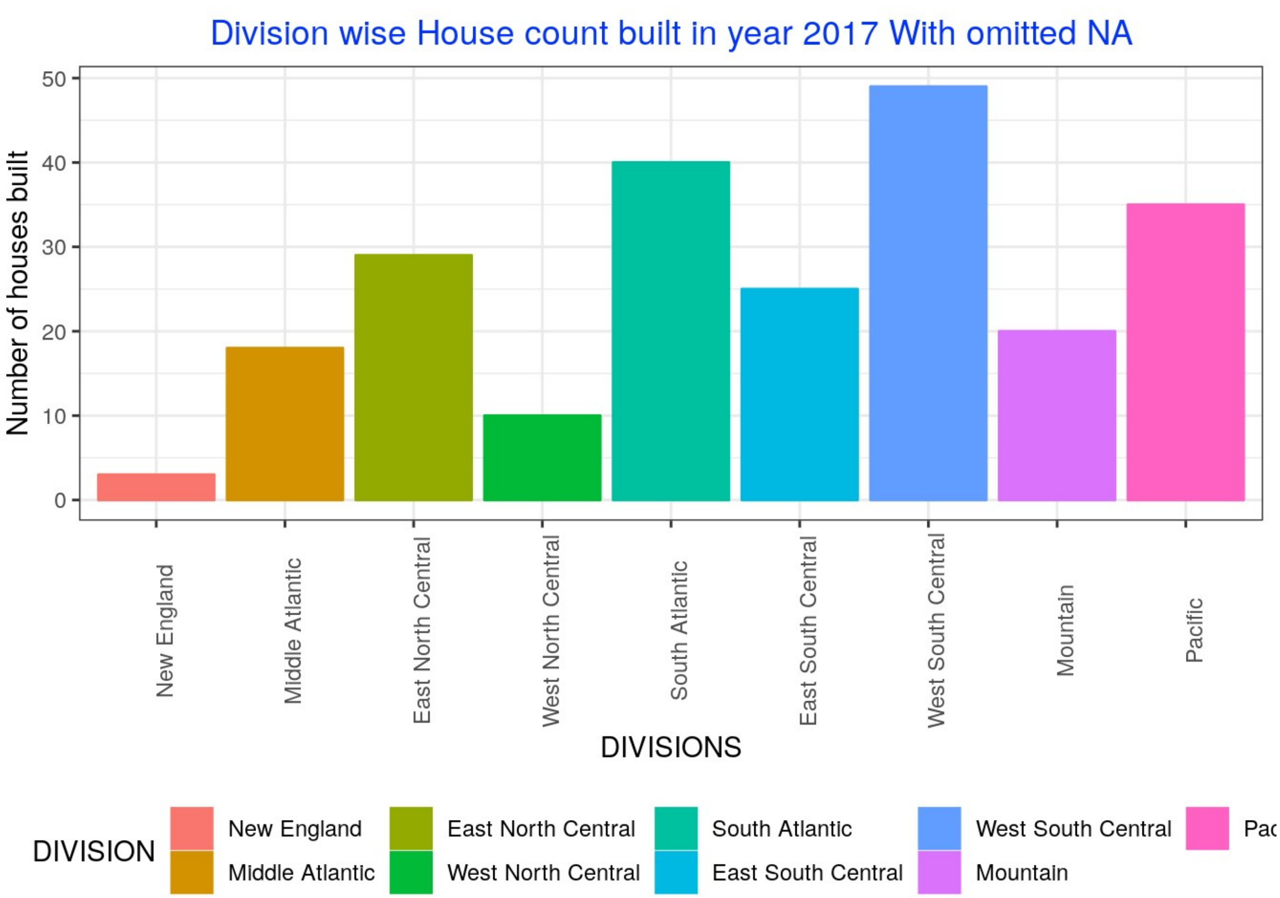

```
#Analysing Divisions Rent in 2017 year by omitting NA
year_division <- Without_NA%>%
  select(DIVISION,Bathtub,RMSP,HouseStructureYear,RENT) %>%
  filter(HouseStructureYear == 2017) %>%
  group_by(DIVISION) %>%
  summarise(Count=n())
year_division
```

DIVISION <fctr>	**Count** <int>
New England	3
Middle Atlantic	18
East North Central	29
West North Central	10
South Atlantic	40
East South Central	25
West South Central	49

DIVISION <fctr>	Count <int>
Mountain	20
Pacific	35

9 rows

```
#plot
ggplot(year_division)+
  geom_col(mapping =aes(x= DIVISION,y= Count,colour =
                        DIVISION,fill=DIVISION))+
  ggtitle("Division wise House count built in year 2017 With omitted NA") +
  xlab("DIVISIONS")+ylab("Number of houses built")+ theme_bw()+
  theme(plot.title= element_text(color="#0033FF",hjust = 0.5),
        axis.text.x = element_text(angle = 90),legend.position =
        "bottom")
```



Rent vs Division in year 2017

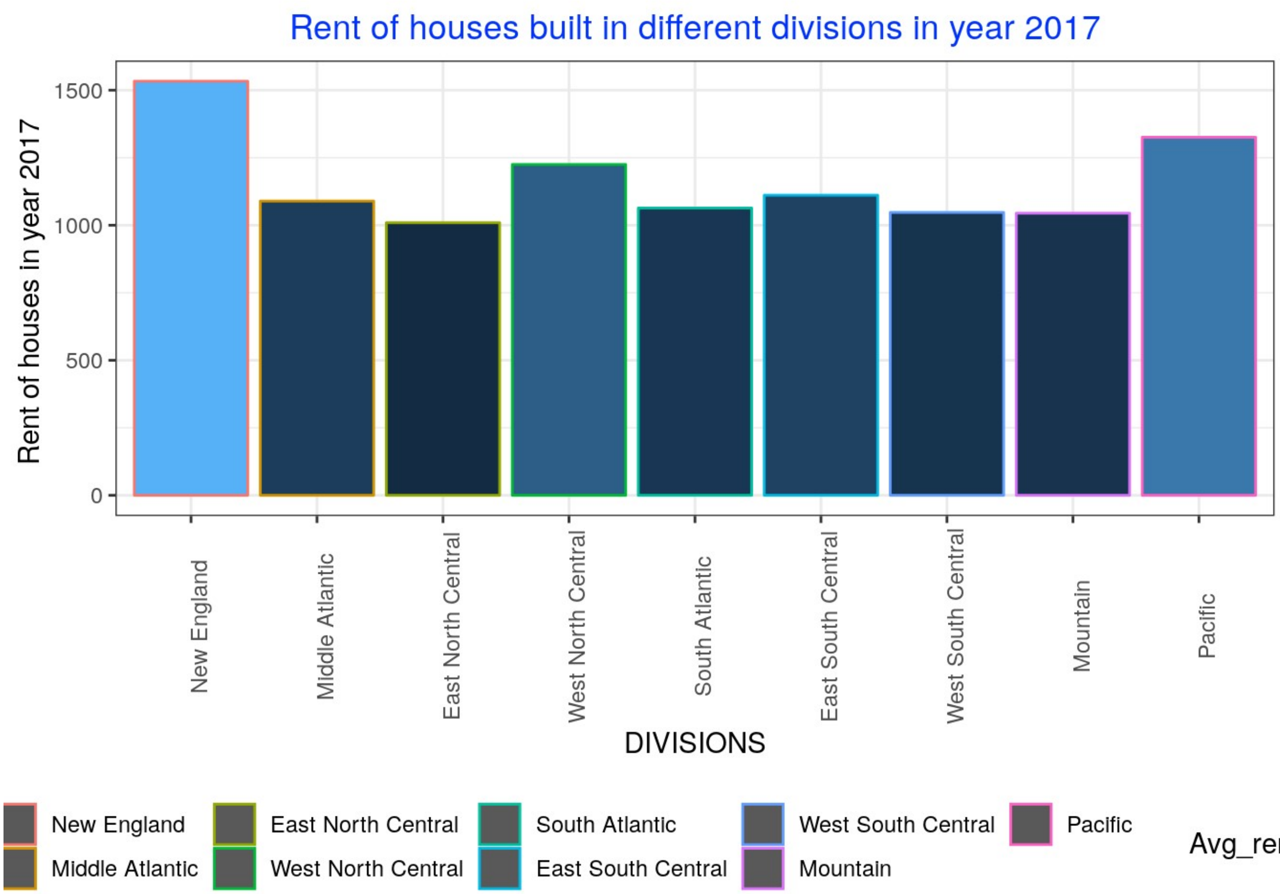
The below graph represents that average Rent of houses built in year 2017 were high in New England and least were in Mountain and East North Central, so investors can easily buy their houses on basis of rent.Moreover, in this graph when I tried with atual data without omitted NA value than I saw that there was no difference, so here I used data with omitted value to visualize data in better way.

```
#Plotting rent ,division wise in year 2017
rent_year <- Without_NA %>%
select(DIVISION,Bathtub,RMSP,HouseStructureYear,RENT) %>%
filter(HouseStructureYear == 2017 ) %>%
group_by(DIVISION) %>% summarise(Avg_rent=mean(RENT))
rent_year
```

DIVISION <fctr>	**Avg_rent** <dbl>
New England	1533.333
Middle Atlantic	1089.444
East North Central	1009.310
West North Central	1225.000
South Atlantic	1063.750
East South Central	1111.200
West South Central	1047.429
Mountain	1044.500
Pacific	1325.829

9 rows

```
#Plot
ggplot(rent_year)+
  geom_col(aes(x= DIVISION,y= Avg_rent,colour = DIVISION,fill=Avg_rent))+
  ggtitle("Rent of houses built in different divisions in year 2017")+
  xlab("DIVISIONS")+ylab("Rent of houses in year 2017 ")+
  theme_bw()+
  theme(plot.title= element_text(color="#0033FF",hjust = 0.5),
        axis.text.x = element_text(angle = 90),
        legend.position = "bottom")
```



Total Rent in different states

From the below graph we can see that Hawaii(HW) of United states consist highest average Rent in comaprison to other states, thus this graph helps buyers to predict that they should not invest in hawaii if their financial statement is quite low, rather than that they should invest their income in west virgina and Arkansas.

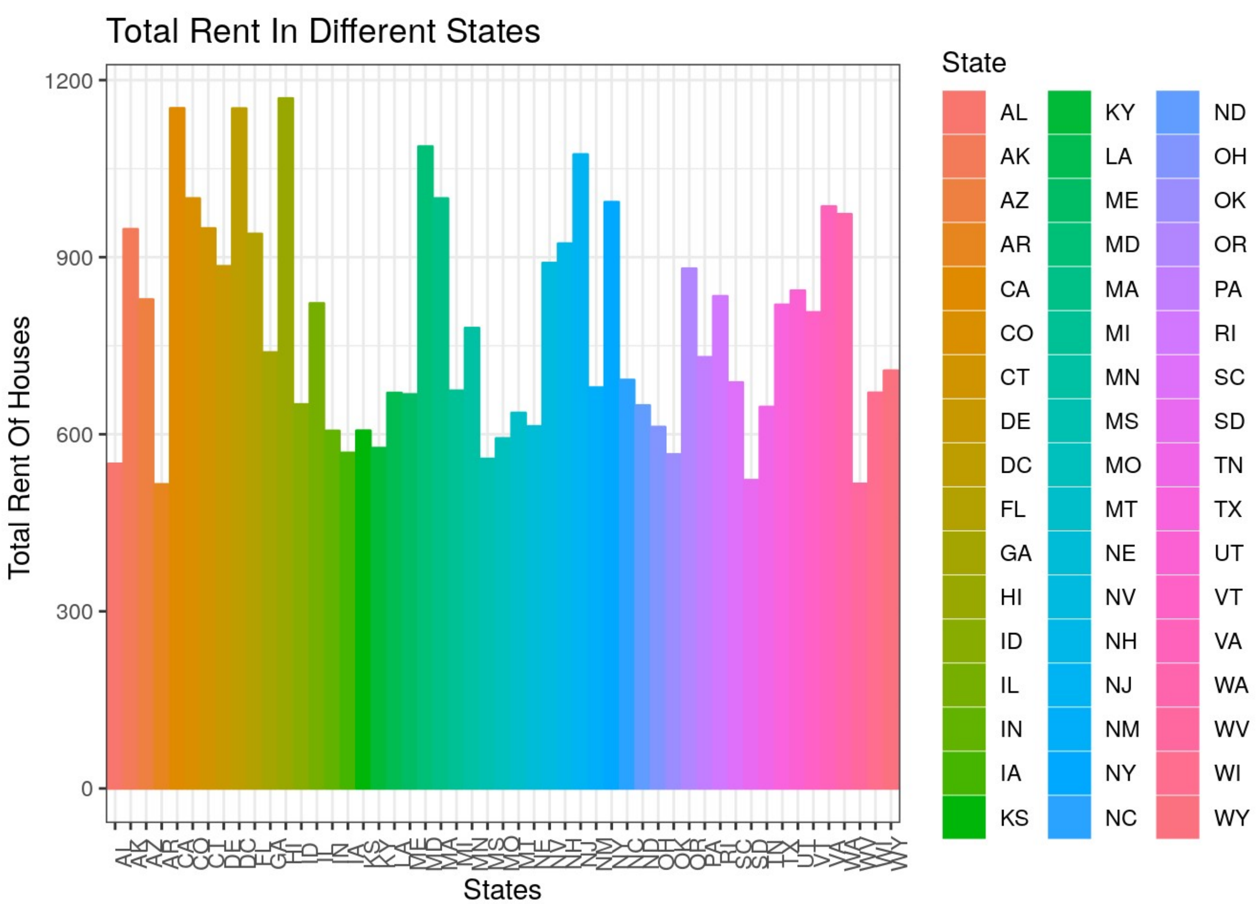
```
#Total rent in different states
rent_rooms <- Without_NA %>%
  select(State,RMSP,HouseStructureYear,RENT)%>%
  group_by(State) %>% summarise(Avg_Rent= mean(RENT))
rent_rooms
```

State <fctr>	Avg_Rent <dbl>
AL	549.8419
AK	947.3652
AZ	828.2409
AR	515.0638
CA	1152.1336

State <fctr>	Avg_Rent <dbl>
CO	999.4572
CT	948.7200
DE	884.5209
DC	1151.8638
FL	939.3696

1-10 of 51 rows Previous **1** 2 3 4 5 6 Next

```
#Plot
ggplot(rent_rooms)+
  geom_col(mapping =aes(x= State,y=Avg_Rent,colour = State,fill=State))+
  ggtitle("Total Rent In Different States")+
  xlab("States")+ylab("Total Rent Of Houses ")+theme_bw()+
  theme(axis.text.x = element_text(angle = 90))
```



Houses built in different years

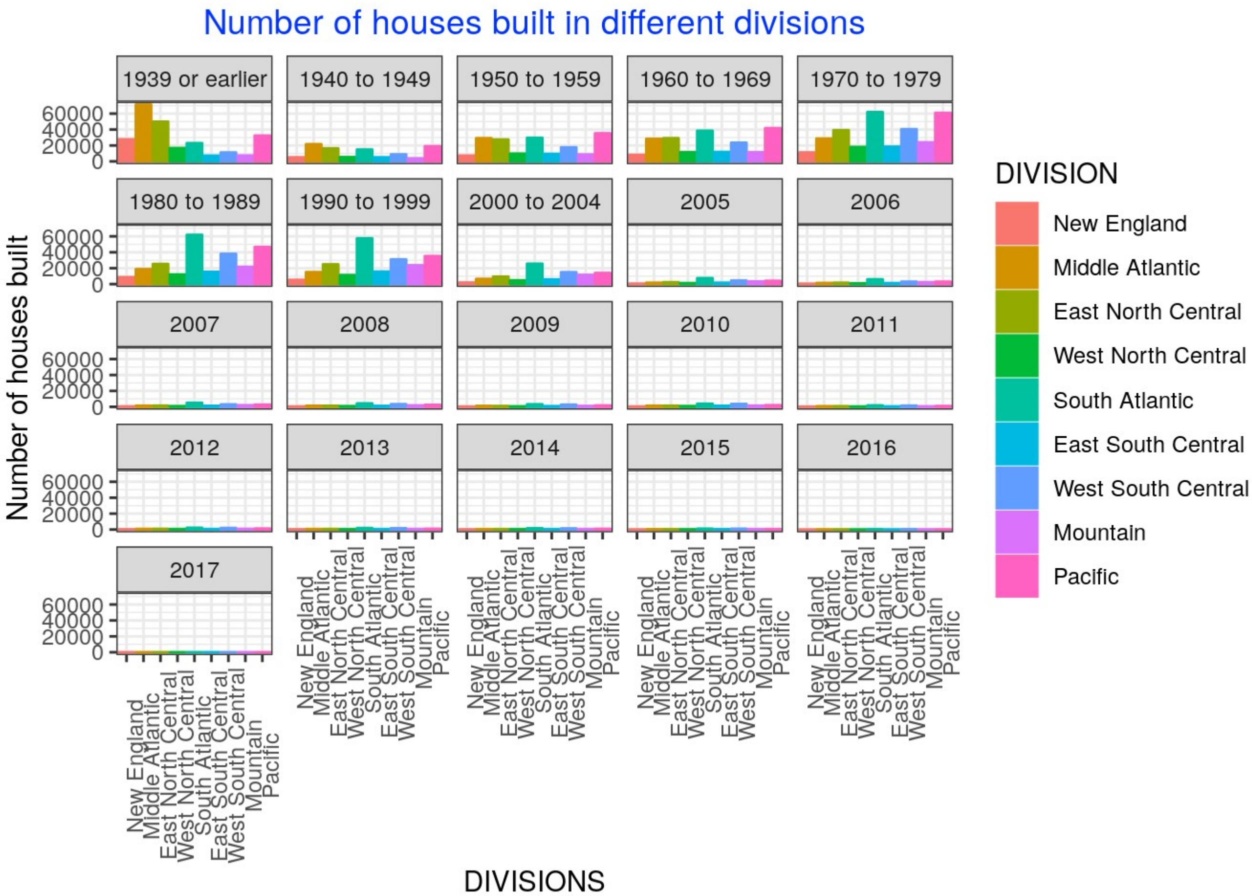
From the below graph, we can say that the ratio of houses built in early years were high compare to 2017, so we can say that houses built in recent years are less.

```
YR_division <- Without_NA%>%
  select(RENT,DIVISION,HouseStructureYear)
YR_division
```

RENT <dbl>	DIVISION <fctr>	HouseStructureYear <fctr>
105.401500	East South Central	1980 to 1989
84.321200	East South Central	1970 to 1979
358.365100	East South Central	2006
1475.621000	East South Central	1990 to 1999
621.868850	East South Central	1980 to 1989
737.810500	East South Central	1950 to 1959
632.409000	East South Central	1980 to 1989
843.212000	East South Central	1980 to 1989
716.730200	East South Central	1980 to 1989
63.240900	East South Central	2000 to 2004

1-10 of 10,000 rows Previous **1** 2 3 4 5 6 ... 1000 Next

```
#plot
options(scipen = 999)
ggplot(YR_division)+geom_bar(mapping =aes(x= DIVISION ,colour =
  DIVISION,fill=DIVISION))+ facet_wrap(~HouseStructureYear)+
  ggtitle("Number of houses built in different divisions")+
  xlab("DIVISIONS")+ylab("Number of houses built")+theme_bw()+
  theme(plot.title= element_text(color="#0033FF",hjust = 0.5),
        axis.text.x = element_text(angle = 90))
```



Tax in different divisions Here from below graph we can see that, tax in South Atlantic is highest and lowest in New England.

So, from overall graphs we can say that it is beneficial to built or rent a house in New England, as it contains lowest price by considering taxes and rent factors.

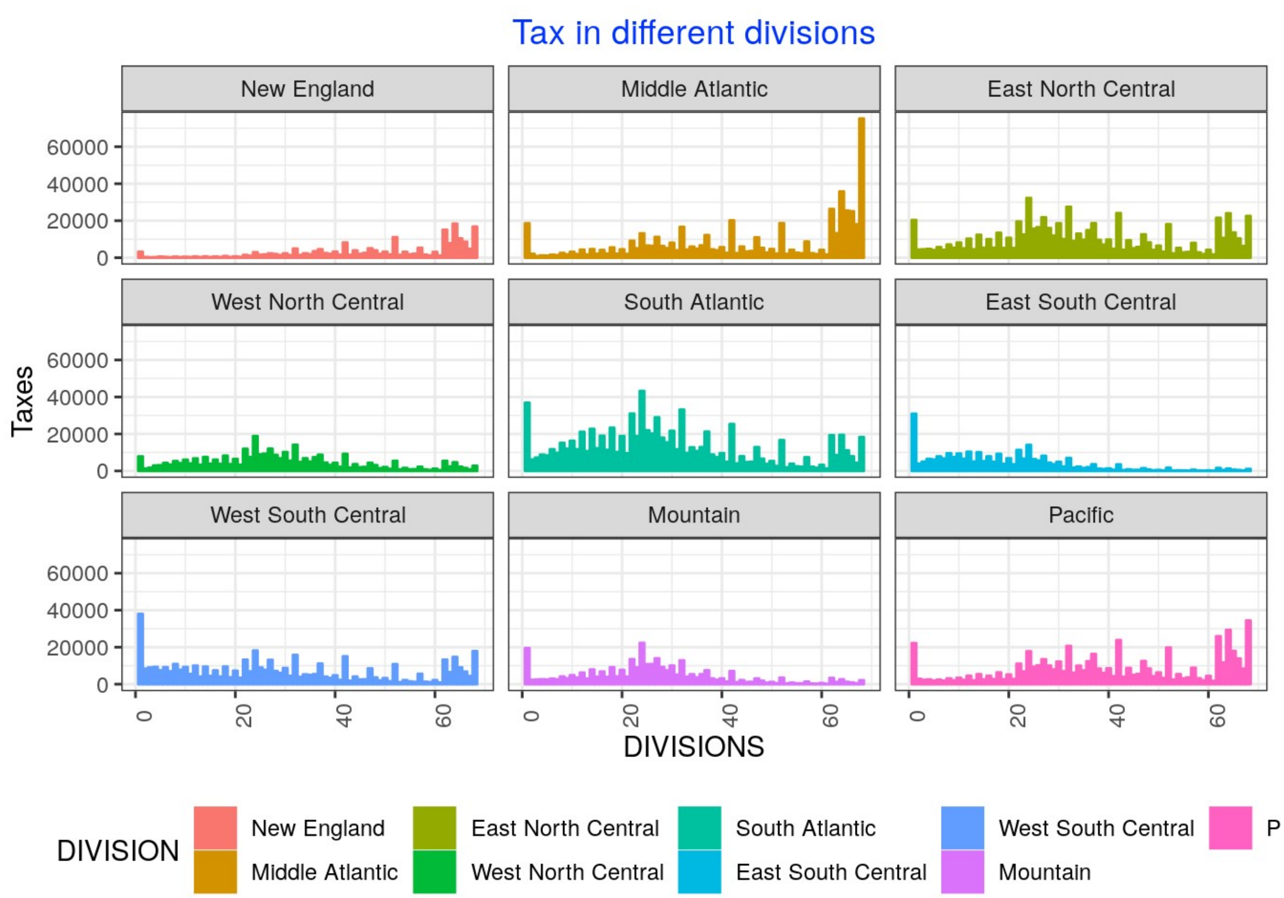
```
tax_division <- bind_data %>%
  select(RENT,DIVISION,HouseStructureYear,Tax)

tax_division
```

RENT <dbl>	DIVISION <fctr>	HouseStructureYear <fctr>	Tax <int>
NA	East South Central	NA	NA
NA	East South Central	1940 to 1949	3
NA	East South Central	1970 to 1979	6
105.40150	East South Central	1980 to 1989	NA
84.32120	East South Central	1970 to 1979	NA
NA	East South Central	1940 to 1949	3

RENT <dbl>	DIVISION <fctr>	HouseStructureYear <fctr>	Tax <int>
NA	East South Central	2000 to 2004	26
NA	East South Central	1940 to 1949	5
358.36510	East South Central	2006	NA
NA	East South Central	1960 to 1969	10

1-10 of 10,000 rows Previous **1** 2 3 4 5 6 ... 1000 Next

```
#plot
options(scipen = 999)
ggplot(tax_division)+geom_bar(mapping =aes(x= Tax ,colour =
  DIVISION,fill=DIVISION))+facet_wrap(~DIVISION)+
  ggtitle("Tax in different divisions")+
  xlab("DIVISIONS")+ylab("Taxes")+theme_bw()+
  theme(plot.title= element_text(color="#0033FF",hjust = 0.5),
        axis.text.x = element_text(angle = 90),legend.position =
        "bottom")
```



Sale of Agriculture product in different division

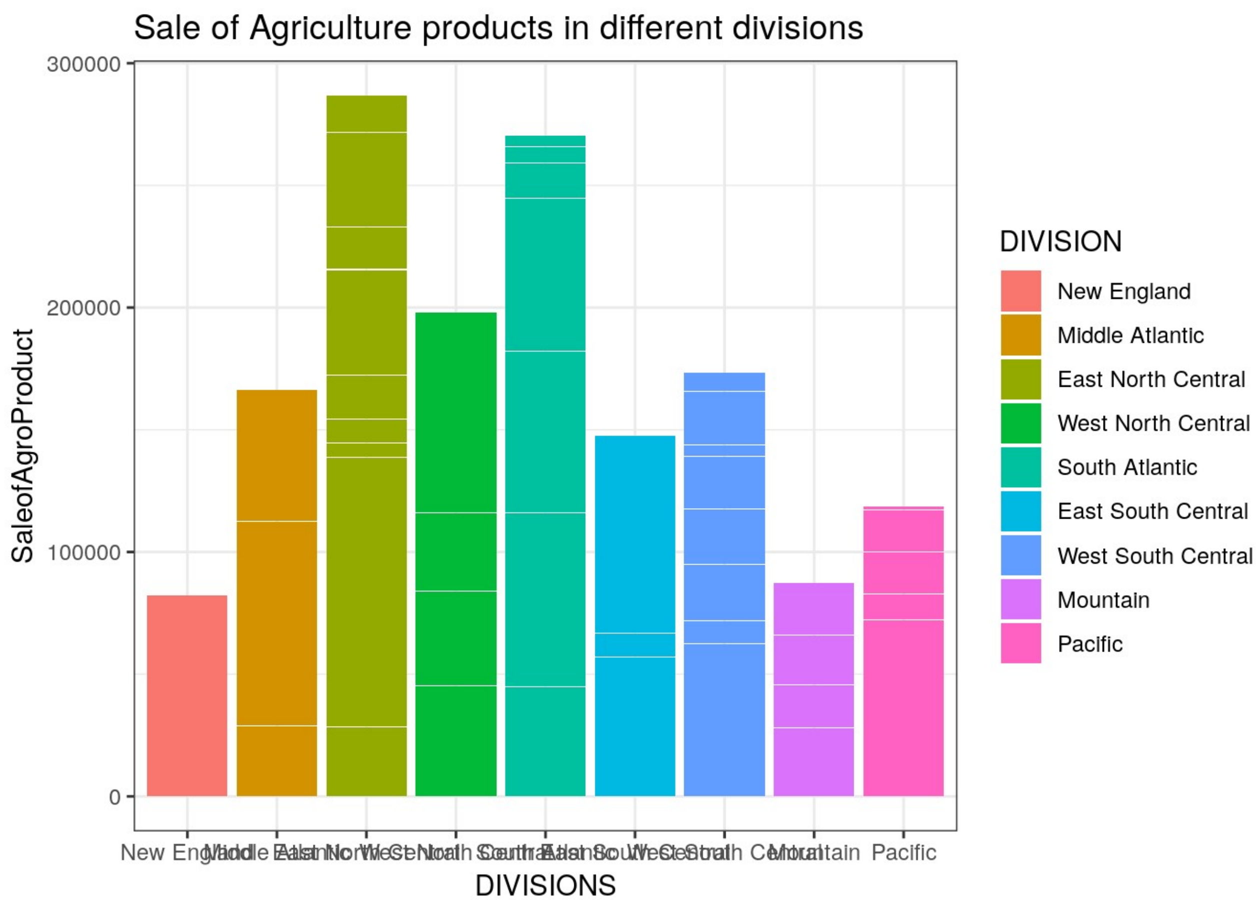
Below graph depicts that East North Central has the highest tax on sale of agriculture products, so if any one wants to do business of agriculture products, they can easily depict from this graph information from where they can get benefit.Moreover, 300000 tax need to pay yearly by East North central which is costly for many individuals.

```
AGS_division <- bind_data %>%
  select(SaleofAgroProduct,DIVISION,Tax) %>% group_by(Tax)

AGS_division
```

SaleofAgroProduct <int>	DIVISION <fctr>	Tax <int>
NA	East South Central	NA
NA	East South Central	3
NA	East South Central	6
1	East South Central	NA
NA	East South Central	NA
NA	East South Central	3
NA	East South Central	26
1	East South Central	5
NA	East South Central	NA
1	East South Central	10

1-10 of 10,000 rows Previous **1** 2 3 4 5 6 ... 1000 Next

```
#plot
ggplot(AGS_division)+
  geom_col(aes(x=DIVISION,y=SaleofAgroProduct,fill=DIVISION))+
  ggtitle("Sale of Agriculture products in different divisions")+
  xlab("DIVISIONS")+ylab("SaleofAgroProduct")+
  theme(axis.text.x = element_text(angle = 90),legend.position = "bottom")+
  theme_bw()
```



Performing Various test for testing P value

From the below performed test we can depict that P value will remain below 0.05, which states that there is significance difference between them, thus it rejects null hypothesis and states that difference of mean of Sale of agro products and tax is not equal to 0 and thus we accept alternative hypothesis.

```
(Variance_test <- var.test(bind_data$SaleofAgroProduct,bind_data$Tax))
```

```
##
##  F test to compare two variances
##
## data:  bind_data$SaleofAgroProduct and bind_data$Tax
## F = 0.002636, num df = 1199657, denom df = 4284627, p-value <
## 0.00000000000000022
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.002629336 0.002642677
## sample estimates:
## ratio of variances
##        0.002635994
```

```
(Variance_test <- var.test(bind_data$HouseAcre,bind_data$Tax))
```

```
##
##  F test to compare two variances
##
## data:  bind_data$HouseAcre and bind_data$Tax
## F = 0.00082329, num df = 5317320, denom df = 4284627, p-value <
## 0.00000000000000022
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.0008221860 0.0008243909
## sample estimates:
## ratio of variances
##       0.0008232881
```

```
(t.test(bind_data$SaleofAgroProduct,bind_data$Tax,data=bind_data))
```

```
##
##  Welch Two Sample t-test
##
## data:  bind_data$SaleofAgroProduct and bind_data$Tax
## t = -3366.7, df = 4364301, p-value < 0.00000000000000022
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -32.63092 -32.59295
## sample estimates:
## mean of x mean of y
##  1.275661 33.887593
```

#Intial model of linear regression for checking AIC and BIC

By checking AIC and BIC we can say that int_model fits best as it has lowest AIC value. From the below image we can say that, the black colour in image means it has not included few variables in that area and the coloured area represents that they are related to log probablity.The log posterior probablity are scaled so 0 represents to lowest probablity from other models.

```
#Intial model of linear regression
int_model <- lm(Tax ~ State+DIVISION+ HouseAcre+ SaleofAgroProduct+
                Bathtub+ HotWater+ Bedrooms+ RMSP+ SINK+ Stove+ Toilet+
                HouseStructureYear+ Kitchen, data = bind_data)
summary(int_model)
```

```
##
## Call:
## lm(formula = Tax ~ State + DIVISION + HouseAcre + SaleofAgroProduct +
##     Bathtub + HotWater + Bedrooms + RMSP + SINK + Stove + Toilet +
##     HouseStructureYear + Kitchen, data = bind_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -85.051  -9.225  -0.782   8.735  75.005
##
## Coefficients: (9 not defined because of singularities)
##                                  Estimate Std. Error t value
## (Intercept)                    -11.978632   0.379278 -31.583
## StateAK                         18.871273   0.322099  58.588
## StateAZ                         16.630868   0.158068 105.213
## StateAR                          5.818015   0.137528  42.304
## StateCA                         30.540127   0.113192 269.808
## StateCO                         15.790298   0.152607 103.470
## StateCT                         47.127950   0.143027 329.504
## StateDE                         10.332911   0.318478  32.445
## StateDC                         20.907547   1.167119  17.914
## StateFL                         18.281802   0.116580 156.818
## StateGA                         13.209778   0.108751 121.468
## StateHI                         16.465641   0.394057  41.785
## StateID                         12.692697   0.193388  65.633
## StateIL                         30.728808   0.122727 250.383
## StateIN                         13.116626   0.124381 105.456
## StateIA                         21.676571   0.159162 136.192
## StateKS                         19.855180   0.167534 118.514
## StateKY                          8.979579   0.127559  70.395
## StateLA                          0.944223   0.139470   6.770
## StateME                         24.142636   0.151454 159.406
## StateMD                         31.233741   0.144384 216.324
## StateMA                         42.071630   0.132679 317.093
## StateMI                         20.482386   0.105622 193.922
## StateMN                         19.033727   0.113131 168.245
## StateMS                          4.828923   0.135862  35.543
## StateMO                         12.753821   0.121243 105.192
## StateMT                         17.648650   0.198275  89.011
## StateNE                         23.154042   0.213143 108.632
## StateNV                         16.530729   0.265788  62.195
## StateNH                         45.060684   0.159561 282.405
## StateNJ                         48.632161   0.146157 332.739
## StateNM                          9.494139   0.189757  50.033
## StateNY                         34.938775   0.106024 329.537
## StateNC                         13.709460   0.108465 126.396
## StateND                          8.127721   0.248994  32.642
## StateOH                         24.089539   0.111405 216.233
```

```
## StateOK                          8.427966   0.135393  62.248
## StateOR                         24.318865   0.156491 155.401
## StatePA                         26.995960   0.104242 258.974
## StateRI                         44.161726   0.304737 144.917
## StateSC                          5.897253   0.126294  46.695
## StateSD                         17.018325   0.249335  68.255
## StateTN                          9.615002   0.112887  85.174
## StateTX                         22.133082   0.102664 215.588
## StateUT                         13.537420   0.225218  60.108
## StateVT                         39.553941   0.195003 202.837
## StateVA                         16.754747   0.117946 142.054
## StateWA                         28.891947   0.132081 218.745
## StateWV                          5.338476   0.170191  31.368
## StateWI                         30.288909   0.109925 275.542
## StateWY                         13.949802   0.298105  46.795
## DIVISIONMiddle Atlantic                NA         NA      NA
## DIVISIONEast North Central             NA         NA      NA
## DIVISIONWest North Central             NA         NA      NA
## DIVISIONSouth Atlantic                 NA         NA      NA
## DIVISIONEast South Central             NA         NA      NA
## DIVISIONWest South Central             NA         NA      NA
## DIVISIONMountain                       NA         NA      NA
## DIVISIONPacific                        NA         NA      NA
## HouseAcre                        0.259257   0.037336   6.944
## SaleofAgroProduct                0.439793   0.015068  29.187
## Bathtub                         -3.311064   0.387065  -8.554
## HotWater                               NA         NA      NA
## Bedrooms                         2.145049   0.018499 115.954
## RMSP                             1.763974   0.007283 242.213
## SINK                             3.099082   0.466792   6.639
## Stove                            5.086451   0.389413  13.062
## Toilet                           0.176973   0.003599  49.177
## HouseStructureYear1940 to 1949   0.915040   0.084415  10.840
## HouseStructureYear1950 to 1959   3.114449   0.067330  46.256
## HouseStructureYear1960 to 1969   3.415948   0.063914  53.446
## HouseStructureYear1970 to 1979   3.425114   0.054510  62.835
## HouseStructureYear1980 to 1989   5.104416   0.055662  91.704
## HouseStructureYear1990 to 1999   6.215056   0.053168 116.895
## HouseStructureYear2000 to 2004   8.684713   0.061854 140.406
## HouseStructureYear2005           9.855163   0.109661  89.870
## HouseStructureYear2006          10.196022   0.112380  90.728
## HouseStructureYear2007          10.493042   0.119794  87.592
## HouseStructureYear2008          10.375215   0.134688  77.031
## HouseStructureYear2009          10.071803   0.160769  62.648
## HouseStructureYear2010           9.775719   0.167285  58.437
## HouseStructureYear2011          10.165844   0.205380  49.498
## HouseStructureYear2012           9.864334   0.199505  49.444
## HouseStructureYear2013          10.777305   0.219550  49.088
## HouseStructureYear2014          10.649474   0.246261  43.245
```

```
## HouseStructureYear2015          10.357083   0.288970  35.841
## HouseStructureYear2016           8.878885   0.414538  21.419
## HouseStructureYear2017           7.995912   0.879378   9.093
## Kitchen                         -7.904206   0.362184 -21.824
##                                             Pr(>|t|)
## (Intercept)                    < 0.0000000000000002 ***
## StateAK                        < 0.0000000000000002 ***
## StateAZ                        < 0.0000000000000002 ***
## StateAR                        < 0.0000000000000002 ***
## StateCA                        < 0.0000000000000002 ***
## StateCO                        < 0.0000000000000002 ***
## StateCT                        < 0.0000000000000002 ***
## StateDE                        < 0.0000000000000002 ***
## StateDC                        < 0.0000000000000002 ***
## StateFL                        < 0.0000000000000002 ***
## StateGA                        < 0.0000000000000002 ***
## StateHI                        < 0.0000000000000002 ***
## StateID                        < 0.0000000000000002 ***
## StateIL                        < 0.0000000000000002 ***
## StateIN                        < 0.0000000000000002 ***
## StateIA                        < 0.0000000000000002 ***
## StateKS                        < 0.0000000000000002 ***
## StateKY                        < 0.0000000000000002 ***
## StateLA                          0.00000000001288 ***
## StateME                        < 0.0000000000000002 ***
## StateMD                        < 0.0000000000000002 ***
## StateMA                        < 0.0000000000000002 ***
## StateMI                        < 0.0000000000000002 ***
## StateMN                        < 0.0000000000000002 ***
## StateMS                        < 0.0000000000000002 ***
## StateMO                        < 0.0000000000000002 ***
## StateMT                        < 0.0000000000000002 ***
## StateNE                        < 0.0000000000000002 ***
## StateNV                        < 0.0000000000000002 ***
## StateNH                        < 0.0000000000000002 ***
## StateNJ                        < 0.0000000000000002 ***
## StateNM                        < 0.0000000000000002 ***
## StateNY                        < 0.0000000000000002 ***
## StateNC                        < 0.0000000000000002 ***
## StateND                        < 0.0000000000000002 ***
## StateOH                        < 0.0000000000000002 ***
## StateOK                        < 0.0000000000000002 ***
## StateOR                        < 0.0000000000000002 ***
## StatePA                        < 0.0000000000000002 ***
## StateRI                        < 0.0000000000000002 ***
## StateSC                        < 0.0000000000000002 ***
## StateSD                        < 0.0000000000000002 ***
## StateTN                        < 0.0000000000000002 ***
## StateTX                        < 0.0000000000000002 ***
```

```
## StateUT                        < 0.0000000000000002 ***
## StateVT                        < 0.0000000000000002 ***
## StateVA                        < 0.0000000000000002 ***
## StateWA                        < 0.0000000000000002 ***
## StateWV                        < 0.0000000000000002 ***
## StateWI                        < 0.0000000000000002 ***
## StateWY                        < 0.0000000000000002 ***
## DIVISIONMiddle Atlantic                         NA
## DIVISIONEast North Central                      NA
## DIVISIONWest North Central                      NA
## DIVISIONSouth Atlantic                          NA
## DIVISIONEast South Central                      NA
## DIVISIONWest South Central                      NA
## DIVISIONMountain                                NA
## DIVISIONPacific                                 NA
## HouseAcre                          0.00000000000382 ***
## SaleofAgroProduct              < 0.0000000000000002 ***
## Bathtub                        < 0.0000000000000002 ***
## HotWater                                        NA
## Bedrooms                       < 0.0000000000000002 ***
## RMSP                           < 0.0000000000000002 ***
## SINK                               0.00000000003157 ***
## Stove                          < 0.0000000000000002 ***
## Toilet                         < 0.0000000000000002 ***
## HouseStructureYear1940 to 1949 < 0.0000000000000002 ***
## HouseStructureYear1950 to 1959 < 0.0000000000000002 ***
## HouseStructureYear1960 to 1969 < 0.0000000000000002 ***
## HouseStructureYear1970 to 1979 < 0.0000000000000002 ***
## HouseStructureYear1980 to 1989 < 0.0000000000000002 ***
## HouseStructureYear1990 to 1999 < 0.0000000000000002 ***
## HouseStructureYear2000 to 2004 < 0.0000000000000002 ***
## HouseStructureYear2005         < 0.0000000000000002 ***
## HouseStructureYear2006         < 0.0000000000000002 ***
## HouseStructureYear2007         < 0.0000000000000002 ***
## HouseStructureYear2008         < 0.0000000000000002 ***
## HouseStructureYear2009         < 0.0000000000000002 ***
## HouseStructureYear2010         < 0.0000000000000002 ***
## HouseStructureYear2011         < 0.0000000000000002 ***
## HouseStructureYear2012         < 0.0000000000000002 ***
## HouseStructureYear2013         < 0.0000000000000002 ***
## HouseStructureYear2014         < 0.0000000000000002 ***
## HouseStructureYear2015         < 0.0000000000000002 ***
## HouseStructureYear2016         < 0.0000000000000002 ***
## HouseStructureYear2017         < 0.0000000000000002 ***
## Kitchen                        < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.54 on 1073351 degrees of freedom
```

```
##   (6413930 observations deleted due to missingness)
## Multiple R-squared:  0.4489, Adjusted R-squared:  0.4489
## F-statistic: 1.107e+04 on 79 and 1073351 DF,  p-value: < 0.00000000000000022
```

```
int_model1 <- lm(Tax ~ Bathtub+HotWater+Bedrooms+RMSP+
                       SINK+Stove+Toilet+HouseStructureYear+
                       Kitchen,data = bind_data)

# Checking which one is better AIC or BIC,Lower the value,
#better the model fits

(aic_model <- AIC(int_model,k=2))
```

```
## [1] 8792634
```

```
(aic_model <- AIC(int_model1,k=2))
```

```
## [1] 37260276
```

```
(bic_model <- BIC(int_model))
```

```
## [1] 8793596
```

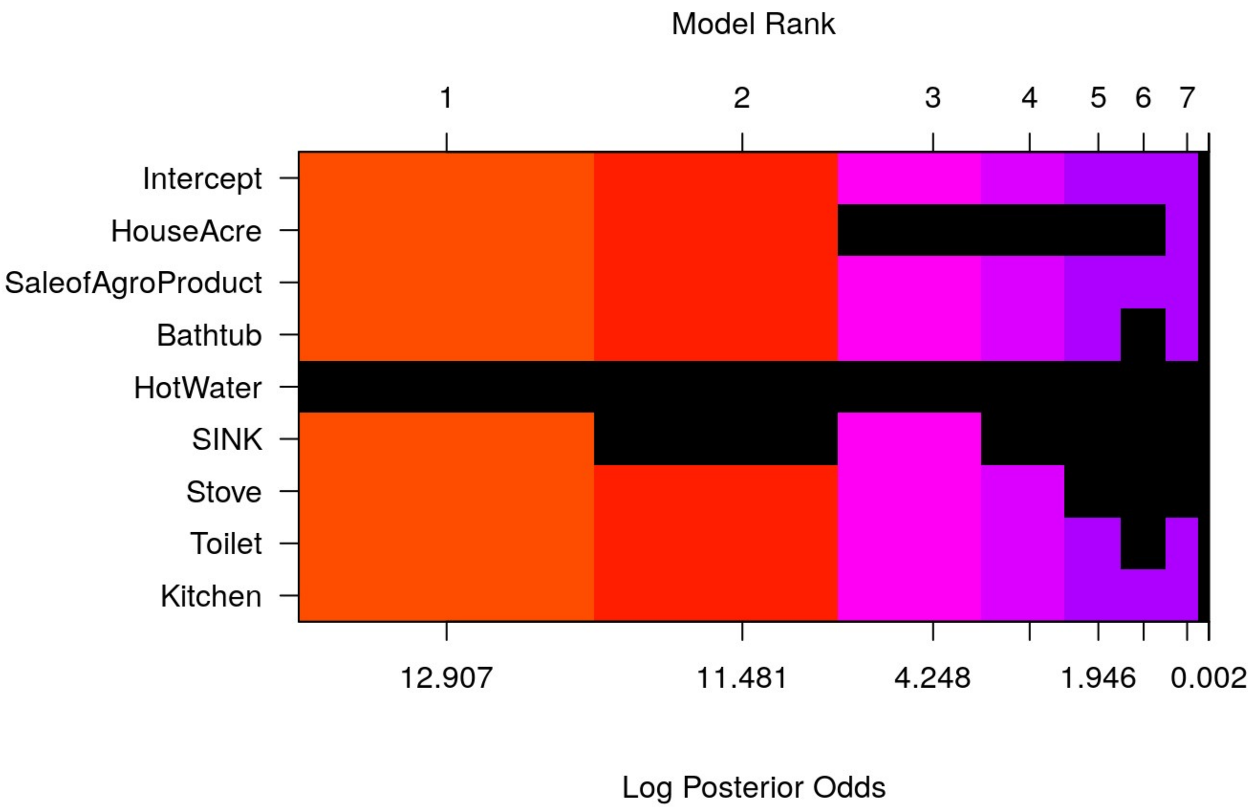
```
(bic_model <- BIC(int_model))
```

```
## [1] 8793596
```

```
#value of AIC model is less so AIC is considered optimal for int_model
model_BAS <- bas.lm(log(Tax) ~ HouseAcre+SaleofAgroProduct
                    +Bathtub+HotWater+SINK+Stove+Toilet+Kitchen,
                    data = bind_data, prior = "AIC", modelprior=uniform(),
                    method = "MCMC", MCMC.iterations=500000)
summary(model_BAS)
```

```
##                    P(B != 0 | Y)       model 1          model 2
## Intercept               1.000000        1.0000        1.0000000
## HouseAcre               0.999748        1.0000        1.0000000
## SaleofAgroProduct       0.999998        1.0000        1.0000000
## Bathtub                 0.999986        1.0000        1.0000000
## HotWater                0.000000        0.0000        0.0000000
## SINK                    0.806194        1.0000        0.0000000
## Stove                   0.999958        1.0000        1.0000000
## Toilet                  0.999984        1.0000        1.0000000
## Kitchen                 1.000000        1.0000        1.0000000
## BF                            NA        1.0000        0.2482373
## PostProbs                     NA        0.8061        0.1937000
## R2                            NA        0.0047        0.0047000
## dim                           NA        8.0000        7.0000000
## logmarg                       NA -7438376.5647 -7438377.9580651
##                            model 3             model 4
## Intercept             1.0000000000       1.00000000000
## HouseAcre             0.0000000000       0.00000000000
## SaleofAgroProduct     1.0000000000       1.00000000000
## Bathtub               1.0000000000       1.00000000000
## HotWater              0.0000000000       0.00000000000
## SINK                  1.0000000000       0.00000000000
## Stove                 1.0000000000       1.00000000000
## Toilet                1.0000000000       1.00000000000
## Kitchen               1.0000000000       1.00000000000
## BF                    0.0003073404       0.00007148497
## PostProbs             0.0001000000       0.00010000000
## R2                    0.0047000000       0.00470000000
## dim                   7.0000000000       6.00000000000
## logmarg           -7438384.6522496780 -7438386.11071831919
##                                   model 5
## Intercept             1.00000000000000000000
## HouseAcre             0.00000000000000000000
## SaleofAgroProduct     1.00000000000000000000
## Bathtub               1.00000000000000000000
## HotWater              0.00000000000000000000
## SINK                  0.00000000000000000000
## Stove                 0.00000000000000000000
## Toilet                1.00000000000000000000
## Kitchen               1.00000000000000000000
## BF                    0.00000000000002643981
## PostProbs             0.00000000000000000000
## R2                    0.00459999999999999992
## dim                   5.00000000000000000000
## logmarg           -7438407.82860064785927534103
```

```
image(model_BAS, rotate = F)
```



Splitting Train and Test data

I am making train data with 75% of train data and rest 25% are test data.

```
set.seed(99)
split <- sample(seq_len(nrow(bind_data)), size = floor(0.75 * nrow(bind_data)))
train <- bind_data[split, ]
test <- bind_data[-split, ]
```

Building Linear Regression Model

I am predicting tax by considering different factors such as state, division, bathtub, sale of agro product and etc, by taking train data. The summary description is explained after result of summary model.

```
model2 <- lm(Tax ~ State+DIVISION+HouseAcre+SaleofAgroProduct+
            Bedrooms+RMSP+HouseStructureYear+SINK+Bathtub+
            Kitchen+INSURANCE, data=train)

(summary(model2))
```

```
##
## Call:
## lm(formula = Tax ~ State + DIVISION + HouseAcre + SaleofAgroProduct +
##     Bedrooms + RMSP + HouseStructureYear + SINK + Bathtub + Kitchen +
##     INSURANCE, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -71.138  -8.379  -0.863   7.662  69.059
##
## Coefficients: (8 not defined because of singularities)
##                                Estimate   Std. Error t value
## (Intercept)                 -12.20285590   0.35251582 -34.616
## StateAK                      17.69582778   0.35623286  49.675
## StateAZ                      17.18290632   0.17311286  99.258
## StateAR                       5.82044765   0.14997625  38.809
## StateCA                      28.01200988   0.12695324 220.648
## StateCO                      13.74604994   0.17356931  79.196
## StateCT                      45.52747842   0.16602066 274.228
## StateDE                      11.70148306   0.34629368  33.791
## StateDC                      19.89317605   1.27809000  15.565
## StateFL                      14.95310148   0.13316932 112.286
## StateGA                      13.16824537   0.11899211 110.665
## StateHI                      14.08507248   0.45394278  31.028
## StateID                      14.19249331   0.20932087  67.803
## StateIL                      30.31831190   0.13395611 226.330
## StateIN                      12.76131462   0.13546385  94.205
## StateIA                      21.88632363   0.17362596 126.054
## StateKS                      18.10230146   0.18915859  95.699
## StateKY                       8.85557620   0.13947921  63.490
## StateLA                      -0.22691249   0.15644730  -1.450
## StateME                      25.16245457   0.16303181 154.341
## StateMD                      31.39316073   0.15967069 196.612
## StateMA                      40.91579165   0.14866974 275.213
## StateMI                      21.15973149   0.11516788 183.729
## StateMN                      17.68805415   0.12521623 141.260
## StateMS                       3.99551931   0.15021201  26.599
## StateMO                      11.81095063   0.13359312  88.410
## StateMT                      17.18363440   0.21975790  78.193
## StateNE                      22.19599338   0.23814842  93.202
## StateNV                      16.99405118   0.29192128  58.214
## StateNH                      45.28557594   0.17287739 261.952
## StateNJ                      48.73236205   0.16476433 295.770
## StateNM                      10.27962047   0.20741884  49.560
## StateNY                      35.05288069   0.11656630 300.712
## StateNC                      14.04629067   0.11845496 118.579
## StateND                       8.50272460   0.27661432  30.739
## StateOH                      24.36428559   0.12160695 200.353
```

```
## StateOK                          6.64013387   0.15198470  43.689
## StateOR                         25.10135562   0.17147528 146.385
## StatePA                         27.77722368   0.11391152 243.849
## StateRI                         42.78573215   0.34929794 122.491
## StateSC                          6.45313242   0.13772622  46.855
## StateSD                         17.40864093   0.27378557  63.585
## StateTN                          9.12465574   0.12352173  73.871
## StateTX                         19.37322916   0.11446258 169.254
## StateUT                         15.58028438   0.24407321  63.834
## StateVT                         39.82245044   0.21145812 188.323
## StateVA                         17.05781070   0.12901229 132.218
## StateWA                         29.05154584   0.14463226 200.865
## StateWV                          7.10160304   0.18373240  38.652
## StateWI                         31.47504954   0.11974143 262.858
## StateWY                         13.10191289   0.33020265  39.678
## DIVISIONMiddle Atlantic                  NA           NA      NA
## DIVISIONEast North Central               NA           NA      NA
## DIVISIONWest North Central               NA           NA      NA
## DIVISIONSouth Atlantic                   NA           NA      NA
## DIVISIONEast South Central               NA           NA      NA
## DIVISIONWest South Central               NA           NA      NA
## DIVISIONMountain                         NA           NA      NA
## DIVISIONPacific                          NA           NA      NA
## HouseAcre                        0.05660256   0.04121198   1.373
## SaleofAgroProduct                0.32641510   0.01715274  19.030
## Bedrooms                         1.43002866   0.02099427  68.115
## RMSP                             1.25823651   0.00848897 148.220
## HouseStructureYear1940 to 1949   1.33113790   0.09201083  14.467
## HouseStructureYear1950 to 1959   3.25372416   0.07376821  44.107
## HouseStructureYear1960 to 1969   3.46099137   0.07010690  49.367
## HouseStructureYear1970 to 1979   3.47062261   0.05983415  58.004
## HouseStructureYear1980 to 1989   4.78196106   0.06136985  77.920
## HouseStructureYear1990 to 1999   5.63575289   0.05873099  95.959
## HouseStructureYear2000 to 2004   7.62648117   0.06871434 110.988
## HouseStructureYear2005           8.86172914   0.12278548  72.172
## HouseStructureYear2006           9.18070061   0.12615696  72.772
## HouseStructureYear2007           9.23974146   0.13471073  68.589
## HouseStructureYear2008           9.12946545   0.15132020  60.332
## HouseStructureYear2009           9.05333219   0.17900000  50.577
## HouseStructureYear2010           9.04870387   0.18662216  48.487
## HouseStructureYear2011           9.41093483   0.22761454  41.346
## HouseStructureYear2012           9.33552999   0.22246062  41.965
## HouseStructureYear2013          10.60784753   0.24356779  43.552
## HouseStructureYear2014          10.84192760   0.27293924  39.723
## HouseStructureYear2015          10.69277643   0.31901977  33.518
## HouseStructureYear2016           8.89603288   0.46102680  19.296
## HouseStructureYear2017           7.10709750   0.99109999   7.171
## SINK                            -0.06339028   0.49832837  -0.127
## Bathtub                         -1.58510938   0.41452801  -3.824
```

```
## Kitchen                          -1.43115857   0.28369996  -5.045
## INSURANCE                         0.00833118   0.00002843 293.064
##                                            Pr(>|t|)
## (Intercept)                     < 0.0000000000000002 ***
## StateAK                         < 0.0000000000000002 ***
## StateAZ                         < 0.0000000000000002 ***
## StateAR                         < 0.0000000000000002 ***
## StateCA                         < 0.0000000000000002 ***
## StateCO                         < 0.0000000000000002 ***
## StateCT                         < 0.0000000000000002 ***
## StateDE                         < 0.0000000000000002 ***
## StateDC                         < 0.0000000000000002 ***
## StateFL                         < 0.0000000000000002 ***
## StateGA                         < 0.0000000000000002 ***
## StateHI                         < 0.0000000000000002 ***
## StateID                         < 0.0000000000000002 ***
## StateIL                         < 0.0000000000000002 ***
## StateIN                         < 0.0000000000000002 ***
## StateIA                         < 0.0000000000000002 ***
## StateKS                         < 0.0000000000000002 ***
## StateKY                         < 0.0000000000000002 ***
## StateLA                                    0.146945
## StateME                         < 0.0000000000000002 ***
## StateMD                         < 0.0000000000000002 ***
## StateMA                         < 0.0000000000000002 ***
## StateMI                         < 0.0000000000000002 ***
## StateMN                         < 0.0000000000000002 ***
## StateMS                         < 0.0000000000000002 ***
## StateMO                         < 0.0000000000000002 ***
## StateMT                         < 0.0000000000000002 ***
## StateNE                         < 0.0000000000000002 ***
## StateNV                         < 0.0000000000000002 ***
## StateNH                         < 0.0000000000000002 ***
## StateNJ                         < 0.0000000000000002 ***
## StateNM                         < 0.0000000000000002 ***
## StateNY                         < 0.0000000000000002 ***
## StateNC                         < 0.0000000000000002 ***
## StateND                         < 0.0000000000000002 ***
## StateOH                         < 0.0000000000000002 ***
## StateOK                         < 0.0000000000000002 ***
## StateOR                         < 0.0000000000000002 ***
## StatePA                         < 0.0000000000000002 ***
## StateRI                         < 0.0000000000000002 ***
## StateSC                         < 0.0000000000000002 ***
## StateSD                         < 0.0000000000000002 ***
## StateTN                         < 0.0000000000000002 ***
## StateTX                         < 0.0000000000000002 ***
## StateUT                         < 0.0000000000000002 ***
## StateVT                         < 0.0000000000000002 ***
```

```
## StateVA                        < 0.0000000000000002 ***
## StateWA                        < 0.0000000000000002 ***
## StateWV                        < 0.0000000000000002 ***
## StateWI                        < 0.0000000000000002 ***
## StateWY                        < 0.0000000000000002 ***
## DIVISIONMiddle Atlantic                         NA    
## DIVISIONEast North Central                      NA    
## DIVISIONWest North Central                      NA    
## DIVISIONSouth Atlantic                          NA    
## DIVISIONEast South Central                      NA    
## DIVISIONWest South Central                      NA    
## DIVISIONMountain                                NA    
## DIVISIONPacific                                 NA    
## HouseAcre                                 0.169613    
## SaleofAgroProduct              < 0.0000000000000002 ***
## Bedrooms                       < 0.0000000000000002 ***
## RMSP                           < 0.0000000000000002 ***
## HouseStructureYear1940 to 1949 < 0.0000000000000002 ***
## HouseStructureYear1950 to 1959 < 0.0000000000000002 ***
## HouseStructureYear1960 to 1969 < 0.0000000000000002 ***
## HouseStructureYear1970 to 1979 < 0.0000000000000002 ***
## HouseStructureYear1980 to 1989 < 0.0000000000000002 ***
## HouseStructureYear1990 to 1999 < 0.0000000000000002 ***
## HouseStructureYear2000 to 2004 < 0.0000000000000002 ***
## HouseStructureYear2005         < 0.0000000000000002 ***
## HouseStructureYear2006         < 0.0000000000000002 ***
## HouseStructureYear2007         < 0.0000000000000002 ***
## HouseStructureYear2008         < 0.0000000000000002 ***
## HouseStructureYear2009         < 0.0000000000000002 ***
## HouseStructureYear2010         < 0.0000000000000002 ***
## HouseStructureYear2011         < 0.0000000000000002 ***
## HouseStructureYear2012         < 0.0000000000000002 ***
## HouseStructureYear2013         < 0.0000000000000002 ***
## HouseStructureYear2014         < 0.0000000000000002 ***
## HouseStructureYear2015         < 0.0000000000000002 ***
## HouseStructureYear2016         < 0.0000000000000002 ***
## HouseStructureYear2017           0.000000000000746 ***
## SINK                                      0.898778    
## Bathtub                                   0.000131 ***
## Kitchen                          0.000000454531502 ***
## INSURANCE                      < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 13.24 on 731544 degrees of freedom
##   (4883897 observations deleted due to missingness)
## Multiple R-squared:  0.5073, Adjusted R-squared:  0.5072 
## F-statistic:  9656 on 78 and 731544 DF,  p-value: < 0.00000000000000022
```

By using train data we can see that accuracy of our model got increased.Moeover, below are some discriptions of summary of model.

Residual Standard Error : It is the average amount that response will deviate from true regression line. In our case actual tax can deviate from true regression line by approximately 13.24.The tax is -12.98 and residual error is 13.24, so our percentage error is 0.26%.

Multiple R-squared : R - squared represents how our model fits the actual data.In our case, the variance is 50% so we can say that some data points will fall near regression and other 50% of data points will be away from regression line. Though we cannot predict exactly that our model will fit our data, however in our case we consider that with 50% of variance we can get predicted model with better accuracy.

Adjusted R- squared : It represents that, as we add on variables into the model, the model gets better and better.

F - statistic : In our case F - statistic value is 9656 is higher than 78, so it suggest that there is relation between predictor and response variable.

```
##Predicting on test data

pred <- predict(model2, newdata=test)

(combine<-data.frame(cbind(test$Tax, pred)))
```

	V1 <dbl>	pred <dbl>
3	6	NA
4	NA	NA
5	NA	NA
8	5	11.99752642
11	NA	NA
14	NA	NA
20	9	NA
24	64	NA
28	1	NA
29	3	2.63236400

1-10 of 10,000 rows Previous **1** 2 3 4 5 6 ... 1000 Next

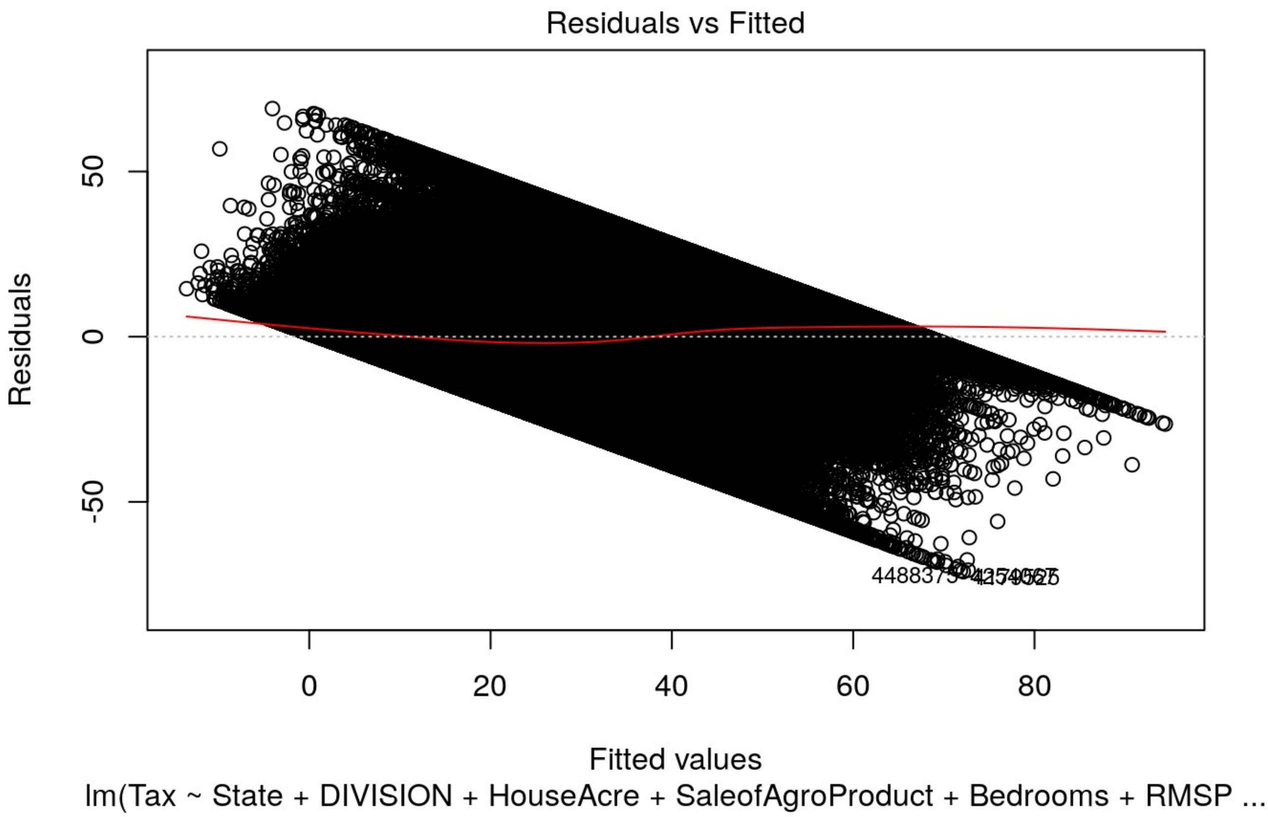
```
colnames(combine)<-c("Actual", "Pred")   # giving column names

(correlation<-cor.test(combine$Actual,combine$Pred))  #correlation
```

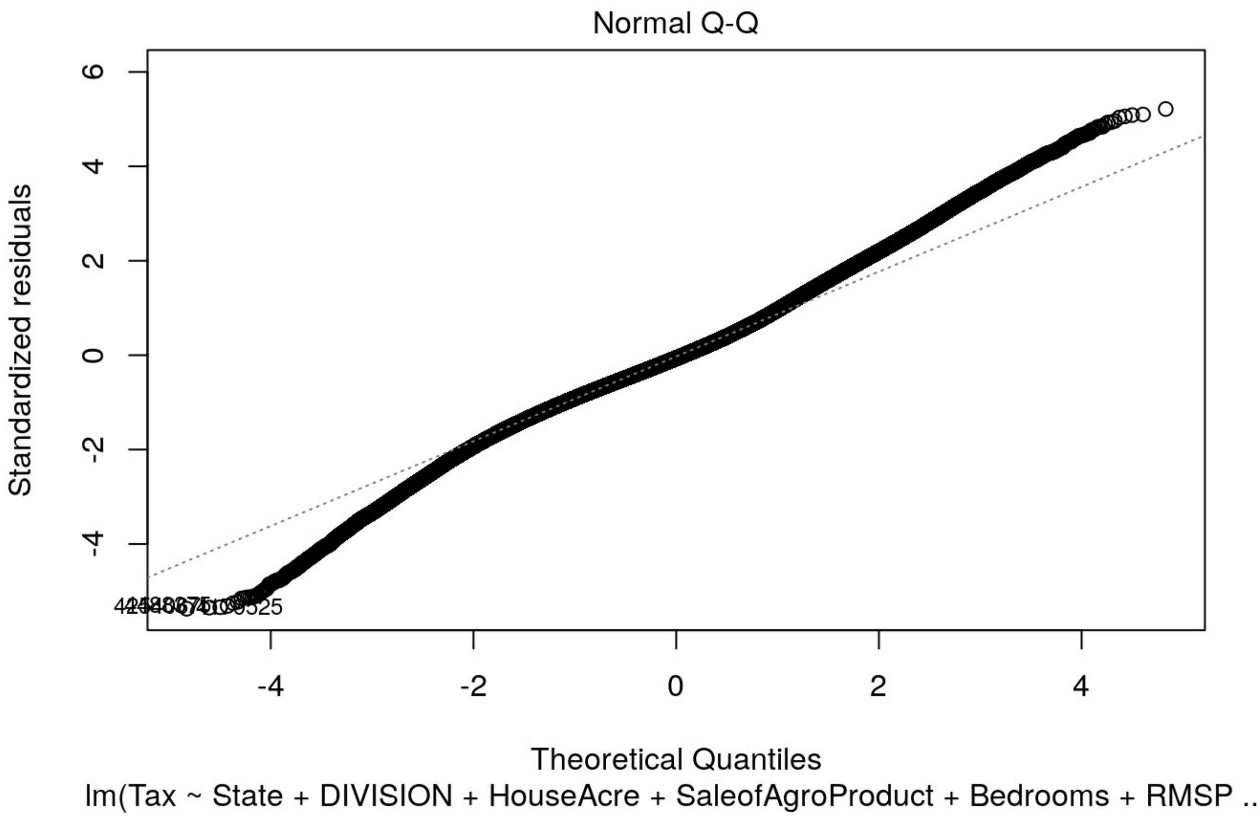
```
##
##  Pearson's product-moment correlation
##
## data:  combine$Actual and combine$Pred
## t = 501.2, df = 244206, p-value < 0.00000000000000022
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7101197 0.7140298
## sample estimates:
##       cor
## 0.7120803
```

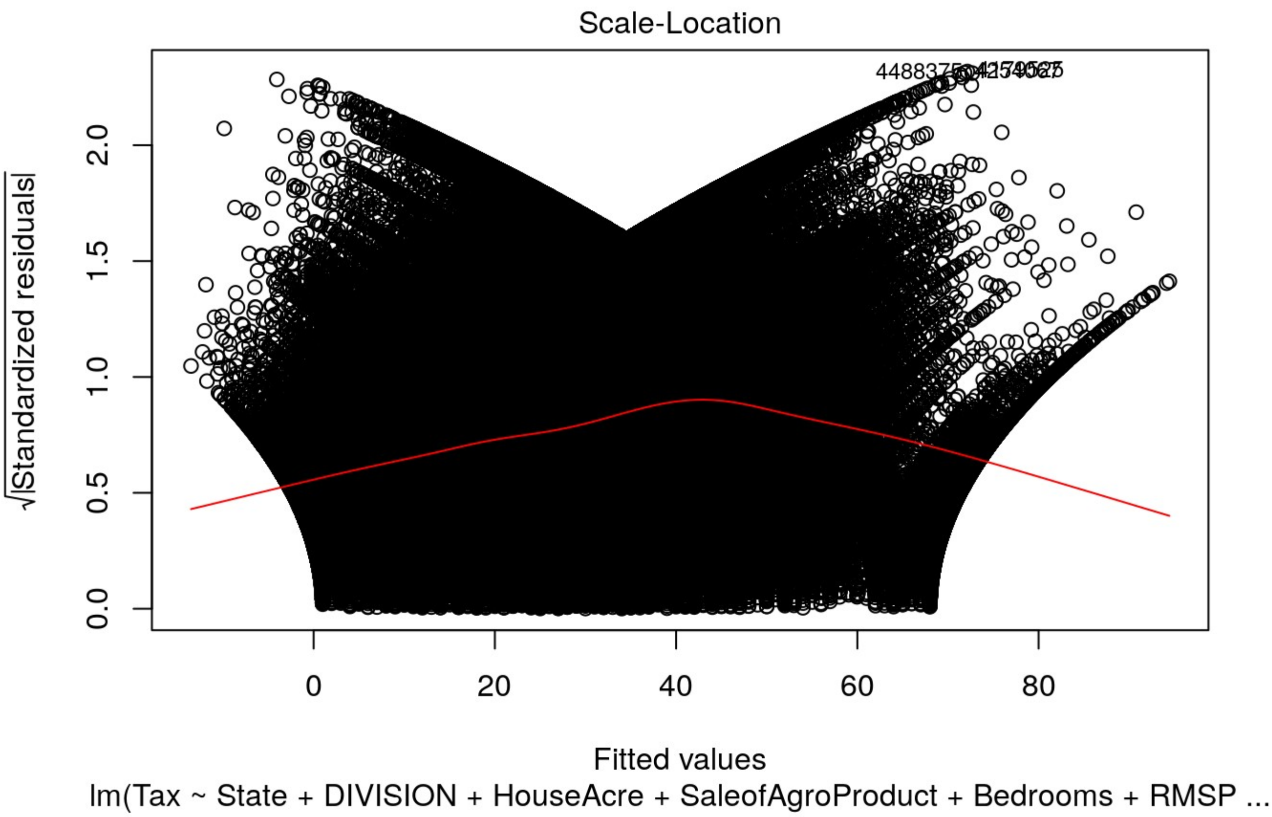
The correlation between Actual and predicted variable is 71%, so it depicts that there is a good relationship between response and predicted variable.Moreover, P - value is less than 0.05 so we reject the Null hypothsesis and we reject that there is relation between tax and other factors. Moreover,For instance, from the combined data frame we can say that on value of actual tax is 5 and predicted tax we got is 11.99.

Plot linear model

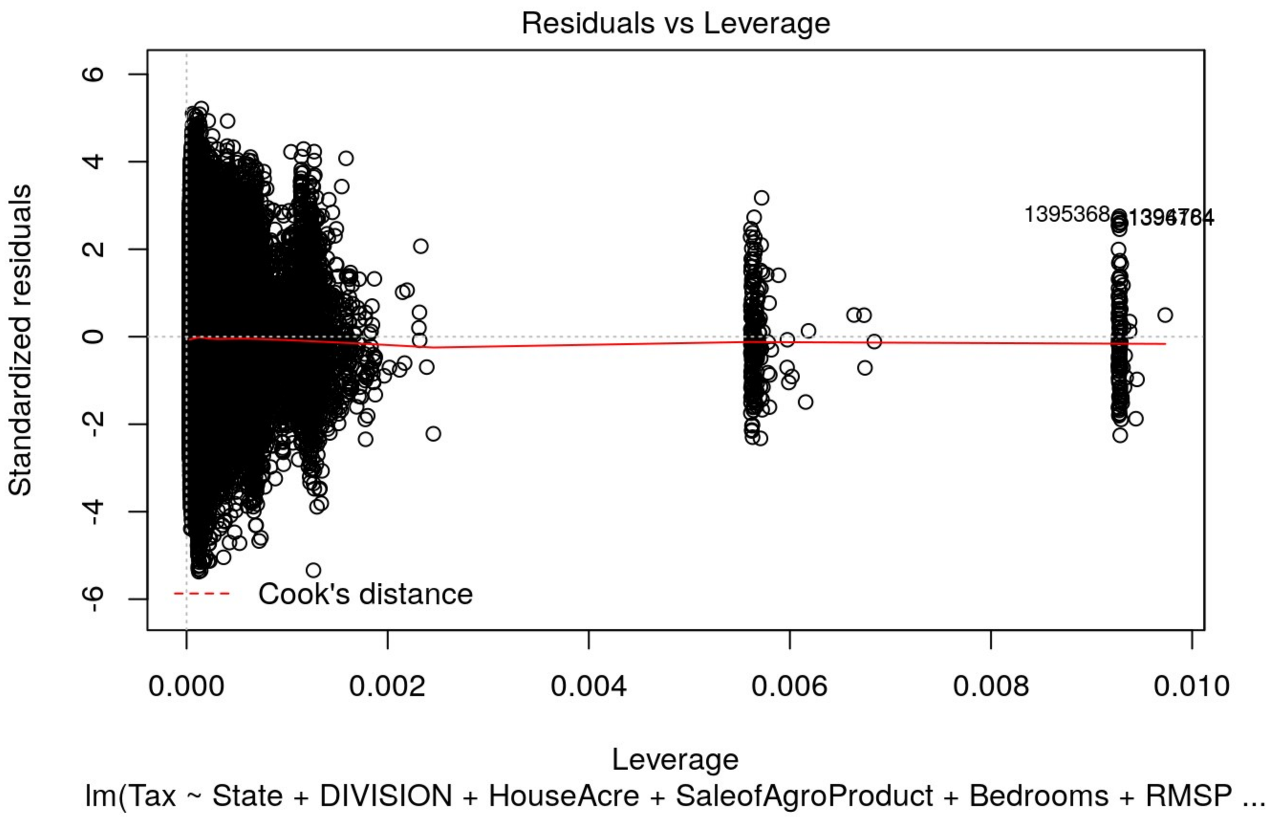
```
plot(model2)
```


Residuals vs Fitted
Residuals
50
0
-50
0
20
40
60
80
Fitted values
lm(Tax ~ State + DIVISION + HouseAcre + SaleofAgroProduct + Bedrooms + RMSP ...


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Tax ~ State + DIVISION + HouseAcre + SaleofAgroProduct + Bedrooms + RMSP ...



Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0
20
40
60
80
Fitted values
lm(Tax ~ State + DIVISION + HouseAcre + SaleofAgroProduct + Bedrooms + RMSP ...



Explanation: Residual vs Fitted graph : From the graph we can see that, as red line shows close relation with dashed line in graph, that means it holds reasonably linearity and also there are some outliners which can affect the model.

Normal Q-Q plot : In this graph, we can see that points fits the centre line well and also there are less ouliers which depicts that Q-Q plot is normally distributed.

Scale location: This graph is used to indicate whether spread of points falls near predicted range or not.So, in our case the residuals shows relation in V shape which means that as red line increases residuals comes near it and as it starts decreasing the points go away from red line.

Residuals vs Leverage: This plot helps us to find influential cases.If the point exceeds from cooks distance that is, from dotted line than it shows that there is high leverage or potential for influencing our model if we exclude that point.In our graph, that is not the case, so we can say that there will be no high influence if we exclude outliers point.

Conclusion:

My main aim was to identify price of house on the basis different utilites, but by performing and analysing some codes I realized that accuracy for that model is so less, in which we cannot predict the actual result. So, to overcome this problem I took linear regression model of tax and other factors and measured the tax on different products. By performing that model I came up with 50% accuracy which was not quite enough for me but as by considering other model with less accuracy, I am quite satisfied with tax linear model.Thus, by analysing model I am somewhat confident with my tax prediction model with 50% accuracy. I dont Know why I am getting less accuracy but this is what I tried and what I got by performing different modelling analysis.I also tried to generate maps on basis of states and division but I could not approach to that level, so I mainly focused on ggplots and plots of linear models.

Appendix

Exra code for variable importance and RMSE check

```
install.packages("Metrics")
library(Metrics)
varImp(model2, scale=FALSE)
```

	Overall <dbl>
StateAK	49.6748889
StateAZ	99.2584037
StateAR	38.8091279
StateCA	220.6482504
StateCO	79.1963161
StateCT	274.2277954
StateDE	33.7906337
StateDC	15.5647693
StateFL	112.2863853
StateGA	110.6648583

1-10 of 78 rows Previous **1** 2 3 4 5 6 ... 8 Next

```
rmse(combine$Actual, combine$Pred)
```

```
## [1] NA
```

I was trying to do linear regression of rent and other factors which can affect the overall price of house but because of less accuracy, I tried to make regression of tax and other factors, from which individuals can predict house on basis of tax in differet divisions, states and other utilities. Below are some code which I tried in making linear regression of Rent including other utilites factors.

```
rent_model <- lm(RENT ~ State+DIVISION+ HouseAcre+ SaleofAgroProduct+
                Bathtub+ HotWater+ Bedrooms+ RMSP+ SINK+ Stove+ Toilet+
                HouseStructureYear+ Kitchen, data = bind_data)
summary(rent_model)
```

```
##
## Call:
## lm(formula = RENT ~ State + DIVISION + HouseAcre + SaleofAgroProduct +
##     Bathtub + HotWater + Bedrooms + RMSP + SINK + Stove + Toilet +
##     HouseStructureYear + Kitchen, data = bind_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1461.56  -220.62   -44.87   169.15  1679.20
##
## Coefficients: (9 not defined because of singularities)
##                               Estimate Std. Error t value
## (Intercept)                   460.2188    29.7227  15.484
## StateAK                       456.4607    22.5039  20.284
## StateAZ                       277.5047    12.8022  21.676
## StateAR                       -21.3679    11.2923  -1.892
## StateCA                       568.0293     8.3560  67.978
## StateCO                       406.1858    12.4908  32.519
## StateCT                       659.9316    14.5749  45.279
## StateDE                       322.5243    25.5864  12.605
## StateDC                       746.6084    86.9272   8.589
## StateFL                       305.9144     9.2477  33.080
## StateGA                       132.0896     8.4571  15.619
## StateHI                       620.6235    22.3202  27.805
## StateID                       151.9215    15.5571   9.765
## StateIL                       197.4104    10.5437  18.723
## StateIN                        93.5840    10.7059   8.741
## StateIA                        -6.4183    13.9916  -0.459
## StateKS                        34.5243    14.1019   2.448
## StateKY                        15.9891    10.5962   1.509
## StateLA                        86.5254    11.9707   7.228
## StateME                       195.9468    13.8302  14.168
## StateMD                       503.2904    12.6861  39.673
## StateMA                       574.0597    13.4204  42.775
## StateMI                       136.0332     9.3690  14.519
## StateMN                       143.4390    10.8887  13.173
## StateMS                        22.9586    11.5369   1.990
## StateMO                        35.8728    10.1059   3.550
## StateMT                       210.4411    15.9646  13.182
## StateNE                         1.4889    16.9086   0.088
## StateNV                       326.2173    17.3829  18.767
## StateNH                       521.6014    15.3580  33.963
## StateNJ                       676.9757    13.7059  49.393
## StateNM                       184.1733    16.4651  11.186
## StateNY                       309.6346     9.2682  33.408
## StateNC                       103.5172     8.5782  12.067
## StateND                        41.4986    27.3215   1.519
## StateOH                       110.4375     9.2568  11.930
```

```
## StateOK                           29.9802    11.2535   2.664
## StateOR                          324.4913    11.5787  28.025
## StatePA                          168.6597     8.9594  18.825
## StateRI                          570.5102    29.9481  19.050
## StateSC                           62.9577    10.1903   6.178
## StateSD                          -52.5596    22.5691  -2.329
## StateTN                           75.9968     9.2880   8.182
## StateTX                          213.0187     8.6059  24.753
## StateUT                          280.0813    18.9938  14.746
## StateVT                          380.8253    17.5987  21.639
## StateVA                          265.7977     9.5149  27.935
## StateWA                          403.6500    10.2859  39.243
## StateWV                           26.5392    16.0326   1.655
## StateWI                          121.5722     9.9197  12.256
## StateWY                          239.0280    26.6035   8.985
## DIVISIONMiddle Atlantic                 NA         NA      NA
## DIVISIONEast North Central              NA         NA      NA
## DIVISIONWest North Central              NA         NA      NA
## DIVISIONSouth Atlantic                  NA         NA      NA
## DIVISIONEast South Central              NA         NA      NA
## DIVISIONWest South Central              NA         NA      NA
## DIVISIONMountain                        NA         NA      NA
## DIVISIONPacific                         NA         NA      NA
## HouseAcre                        -74.4863     3.2002 -23.276
## SaleofAgroProduct                -10.5516     1.4556  -7.249
## Bathtub                         -135.6574    27.6296  -4.910
## HotWater                                NA         NA      NA
## Bedrooms                          52.3548     1.6277  32.165
## RMSP                              25.9464     0.7963  32.585
## SINK                              75.0425    33.2013   2.260
## Stove                              7.2099    26.5461   0.272
## Toilet                             1.8393     0.2988   6.155
## HouseStructureYear1940 to 1949   14.3280     5.4111   2.648
## HouseStructureYear1950 to 1959   61.7424     4.6436  13.296
## HouseStructureYear1960 to 1969   73.5644     4.6455  15.836
## HouseStructureYear1970 to 1979   72.3659     4.1849  17.292
## HouseStructureYear1980 to 1989   81.6534     4.4012  18.552
## HouseStructureYear1990 to 1999   92.0277     4.3701  21.059
## HouseStructureYear2000 to 2004  128.6929     5.9223  21.730
## HouseStructureYear2005          188.1542    11.2695  16.696
## HouseStructureYear2006          224.8361    11.7077  19.204
## HouseStructureYear2007          199.1017    12.8633  15.478
## HouseStructureYear2008          202.3871    14.2476  14.205
## HouseStructureYear2009          206.7902    17.2536  11.985
## HouseStructureYear2010          194.6655    16.0304  12.144
## HouseStructureYear2011          121.9762    22.8231   5.344
## HouseStructureYear2012          202.4161    22.2270   9.107
## HouseStructureYear2013          195.7962    26.9244   7.272
## HouseStructureYear2014          184.1997    32.4892   5.670
```

```
## HouseStructureYear2015           215.4776    37.7444   5.709
## HouseStructureYear2016           188.4295    52.9710   3.557
## HouseStructureYear2017           -15.3768   173.4248  -0.089
## Kitchen                         -109.1227    24.3165  -4.488
##                                             Pr(>|t|)
## (Intercept)                     < 0.0000000000000002 ***
## StateAK                         < 0.0000000000000002 ***
## StateAZ                         < 0.0000000000000002 ***
## StateAR                                     0.058461 .
## StateCA                         < 0.0000000000000002 ***
## StateCO                         < 0.0000000000000002 ***
## StateCT                         < 0.0000000000000002 ***
## StateDE                         < 0.0000000000000002 ***
## StateDC                         < 0.0000000000000002 ***
## StateFL                         < 0.0000000000000002 ***
## StateGA                         < 0.0000000000000002 ***
## StateHI                         < 0.0000000000000002 ***
## StateID                         < 0.0000000000000002 ***
## StateIL                         < 0.0000000000000002 ***
## StateIN                         < 0.0000000000000002 ***
## StateIA                                     0.646434
## StateKS                                     0.014359 *
## StateKY                                     0.131317
## StateLA                         0.000000000000493802 ***
## StateME                         < 0.0000000000000002 ***
## StateMD                         < 0.0000000000000002 ***
## StateMA                         < 0.0000000000000002 ***
## StateMI                         < 0.0000000000000002 ***
## StateMN                         < 0.0000000000000002 ***
## StateMS                                     0.046592 *
## StateMO                                     0.000386 ***
## StateMT                         < 0.0000000000000002 ***
## StateNE                                     0.929835
## StateNV                         < 0.0000000000000002 ***
## StateNH                         < 0.0000000000000002 ***
## StateNJ                         < 0.0000000000000002 ***
## StateNM                         < 0.0000000000000002 ***
## StateNY                         < 0.0000000000000002 ***
## StateNC                         < 0.0000000000000002 ***
## StateND                                     0.128791
## StateOH                         < 0.0000000000000002 ***
## StateOK                                     0.007722 **
## StateOR                         < 0.0000000000000002 ***
## StatePA                         < 0.0000000000000002 ***
## StateRI                         < 0.0000000000000002 ***
## StateSC                         0.000000000651281176 ***
## StateSD                                     0.019871 *
## StateTN                         0.000000000000000282 ***
## StateTX                         < 0.0000000000000002 ***
```

```
## StateUT                         < 0.0000000000000002 ***
## StateVT                         < 0.0000000000000002 ***
## StateVA                         < 0.0000000000000002 ***
## StateWA                         < 0.0000000000000002 ***
## StateWV                                     0.097863 .
## StateWI                         < 0.0000000000000002 ***
## StateWY                         < 0.0000000000000002 ***
## DIVISIONMiddle Atlantic                           NA
## DIVISIONEast North Central                        NA
## DIVISIONWest North Central                        NA
## DIVISIONSouth Atlantic                            NA
## DIVISIONEast South Central                        NA
## DIVISIONWest South Central                        NA
## DIVISIONMountain                                  NA
## DIVISIONPacific                                   NA
## HouseAcre                       < 0.0000000000000002 ***
## SaleofAgroProduct               0.000000000000423541 ***
## Bathtub                         0.000000913056005648 ***
## HotWater                                          NA
## Bedrooms                        < 0.0000000000000002 ***
## RMSP                            < 0.0000000000000002 ***
## SINK                                        0.023809 *
## Stove                                       0.785930
## Toilet                          0.000000000752101470 ***
## HouseStructureYear1940 to 1949              0.008101 **
## HouseStructureYear1950 to 1959 < 0.0000000000000002 ***
## HouseStructureYear1960 to 1969 < 0.0000000000000002 ***
## HouseStructureYear1970 to 1979 < 0.0000000000000002 ***
## HouseStructureYear1980 to 1989 < 0.0000000000000002 ***
## HouseStructureYear1990 to 1999 < 0.0000000000000002 ***
## HouseStructureYear2000 to 2004 < 0.0000000000000002 ***
## HouseStructureYear2005         < 0.0000000000000002 ***
## HouseStructureYear2006         < 0.0000000000000002 ***
## HouseStructureYear2007         < 0.0000000000000002 ***
## HouseStructureYear2008         < 0.0000000000000002 ***
## HouseStructureYear2009         < 0.0000000000000002 ***
## HouseStructureYear2010         < 0.0000000000000002 ***
## HouseStructureYear2011         0.000000090935580362 ***
## HouseStructureYear2012         < 0.0000000000000002 ***
## HouseStructureYear2013         0.000000000000356943 ***
## HouseStructureYear2014         0.000000014359983419 ***
## HouseStructureYear2015         0.000000011409434055 ***
## HouseStructureYear2016                     0.000375 ***
## HouseStructureYear2017                     0.929348
## Kitchen                        0.000007212301045274 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 346.5 on 88829 degrees of freedom
```

```
##   (7398452 observations deleted due to missingness)
## Multiple R-squared:  0.2485, Adjusted R-squared:  0.2478
## F-statistic: 371.8 on 79 and 88829 DF,  p-value: < 0.00000000000000022
```

```
p2 <- predict(rent_model, newdata=test)

(combine<-data.frame(cbind(test$RENT, p2)))
```

	V1 <dbl>	p2 <dbl>
3	NA	NA
4	105.40150	512.2706
5	84.32120	NA
8	NA	441.1304
11	NA	NA
14	621.86885	NA
20	NA	NA
24	NA	934.8959
28	NA	NA
29	NA	544.7764

1-10 of 10,000 rows Previous **1** 2 3 4 5 6 ... 1000 Next

```
colnames(combine)<-c("Actual", "Pred")   # giving column names

(correlation<-cor.test(combine$Actual,combine$Pred))
```

```
##
##  Pearson's product-moment correlation
##
## data:  combine$Actual and combine$Pred
## t = 86.086, df = 22222, p-value < 0.00000000000000022
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4901643 0.5098840
## sample estimates:
##       cor
## 0.500089
```

The accuracy of Rent model is only 24% and from prediction we can say that by including other utilities such as Bathtub, kitchen, agro products, acres, the rent would be 105.40 whilst our predicted rent is 512, which shows a great difference between actual and predicted value. Thus, lower the accuracy, more worst our prediction would be.