Connecticut Rental Housing Market Analysis

Predicting Monthly Rental Prices from Census Housing Data

Mihir Sanghvi

11/23/16

[Course title]

Contents

Introduction

I have analyzed Connecticut rental housing prices (factors affecting the rent) with the help of publicly available PUMS (Public Use Microdata Samples) dataset from census.gov website.

Problem Definition

Predicting Rental Property Prices based on the Rental Property Characteristics using US Census Housing Data.

Dataset Source

http://www.census.gov/programs-surveys/acs/data/pums.html

Note: Downloaded Connecticut Housing Unit Records (Go to 2015 ACS 1-year PUMS) Dataset from the above-mentioned link.

Data Dictionary Source

http://www2.census.gov/programs-surveys/acs/tech_docs/pums/data_dict/PUMSDataDict15.pdf

This dataset provides household level information about the social, economic and financial characteristics of the population of various states, Analysis was done on the data of Connecticut. The Columns and the data are explained in the appendix Data Dictionary.

Initial Data Screening

In the data dictionary, it was clearly mentioned what is the meaning of missing value for all the variables.

i.e.

Variable Name	Levels	Description
ACR (Lot Size)	N/A	GQ / Not a one-family house or mobile home
	1	House on less than one acre
	2	House on one to less than ten acres
	3	House on ten or more acres

It means that if the data is missing in above column then it could be due to its housing type - Group Quarters or if that is not a single-family house or mobile home.

I have analyzed all the variables which could be used to filter the housing data to get rental housing data. I have identified total filter 4 variables as follows:

1. NP - Vacant Unit (0) [Remove All 0s]
2. WGTP - Group Quarters -GQ (0) [Remove All 0s]
3. TYPE - Housing Unit (1) [Keep only 1s]

4. TEN - Rented (3) [Keep Only 3s]

In the dataset, there were 235 variables, I have gone through variable description and identified 20 potential predictors for the given business problem.

Data Cleaning

After Identifying key features with the help of data dictionary, I have performed data cleaning.

There were some variables coded as follows:

Variable Name	Levels	Description
BATH (Bathtub or Shower)	N/A	Group Quarters
	1	Yes
	2	No

Since I have already removed all NA's in this case (by filtering out Group Quarters data) I have decided to recode the variable as follows:

Variable Name	Levels	Description
BATH (Bathtub or shower)	1	House on less than one acre
	2	House on one to less than ten acres

Missing Value Treatment

I have checked missing values in the rental housing dataset and found that 2 columns had same number of missing values and after checking the data dictionary I found that it was due to the fact that that housing record belongs to "not a one-family house or mobile home".

Variable Name	Levels	Description
BUS (Business or medical office on property)	N/A	GQ / Not a one-family house or mobile home
	1	Yes
	2	No
ACR (Lot Size)	N/A	GQ / Not a one-family house or mobile home
	1	House on less than one acre
	2	House on one to less than ten acres
	3	House on ten or more acres

I have recoded above columns as follows:

Variable Name	Levels	Description
BUS (Business or medical office on property)	1	Yes
	2	No
	3	Not a one-family house or mobile home

ACR (Lot Size)	1	House on less than one acre
	2	House on one to less than ten acres
	3	House on ten or more acres
	4	Not a one-family house or mobile home

Feature Engineering

Found that some of the variables were coded in a way where we cannot directly use it in the modeling.

Variable Name	Levels	Description
ELEP (Monthly Cost)	N/A	GQ / Vacant
	1	Included in rent or in condo fee
	2	No charge or electricity not used
	3 - 999	$3 to $999 (Rounded and top-coded)

As we can see above,

- 1 refers to the case when electricity cost is added into monthly rent or condo fee
- 2 refers to the case when no electricity cost is taken from tenants/ not used
- 3-999 refers to the actual monthly cost (in dollars) of electricity
- Note: We don't need to worry about N/A cases because we have already filtered out all GQ (Group Quarters) and Vacant household data

I have created dummy variables for above mentioned cases as follows:

Variable Name	Levels	Description
ELEPIR (Electricity included in rent or in condo fee)	0	No
	1	Yes

Variable Name	Levels	Description
ELEPNC (No charge or electricity not used)	0	No
	1	Yes

After creating the above-mentioned dummy variables, I have recoded ELEP variable as follows

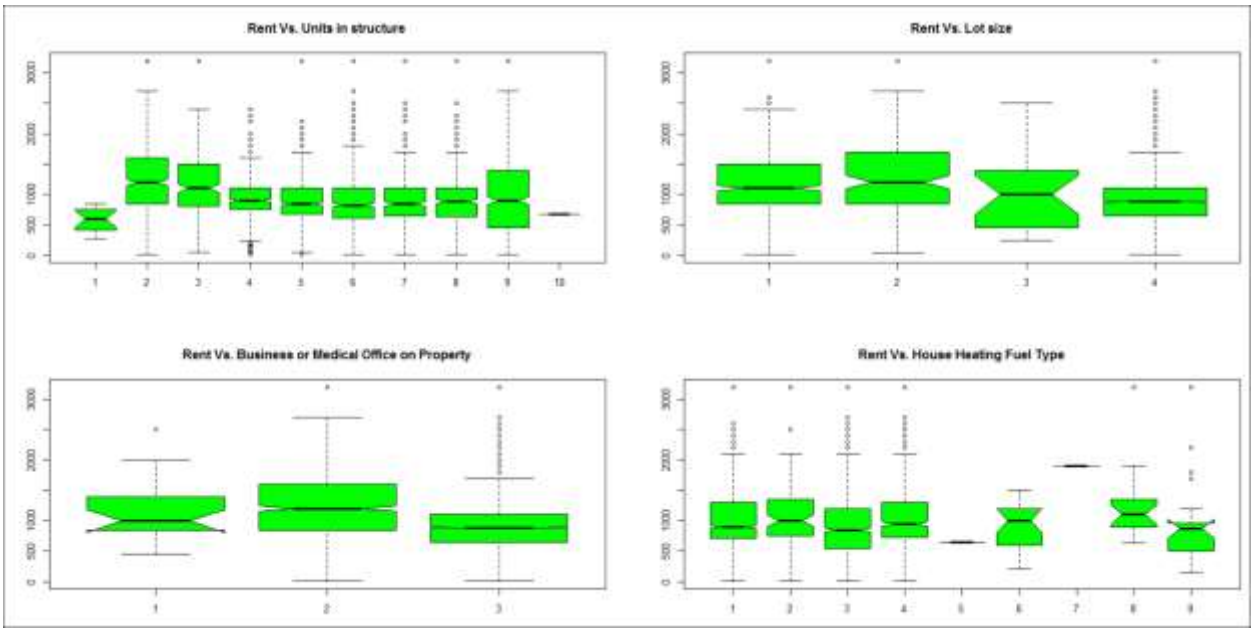
Variable Name	Levels	Description
ELEP (Monthly Cost)	$0	No charge or electricity not used OR Included in rent or in condo fee
	$3 - $999	Monthly Electricity Cost

Similarly, I have created dummy variables for the following variables

1. FULP
2. GASP
3. WATP

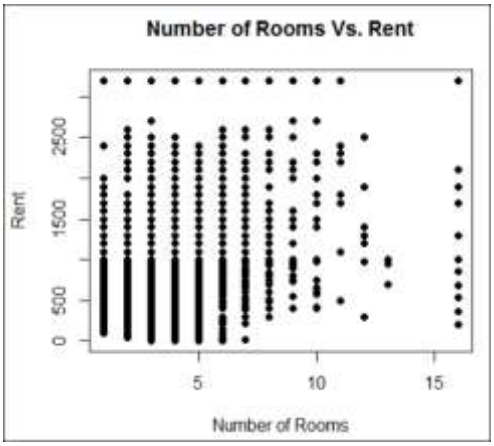
Data Visualization

Visualization #1: Categorical variables with Rent Amount - Box Plot



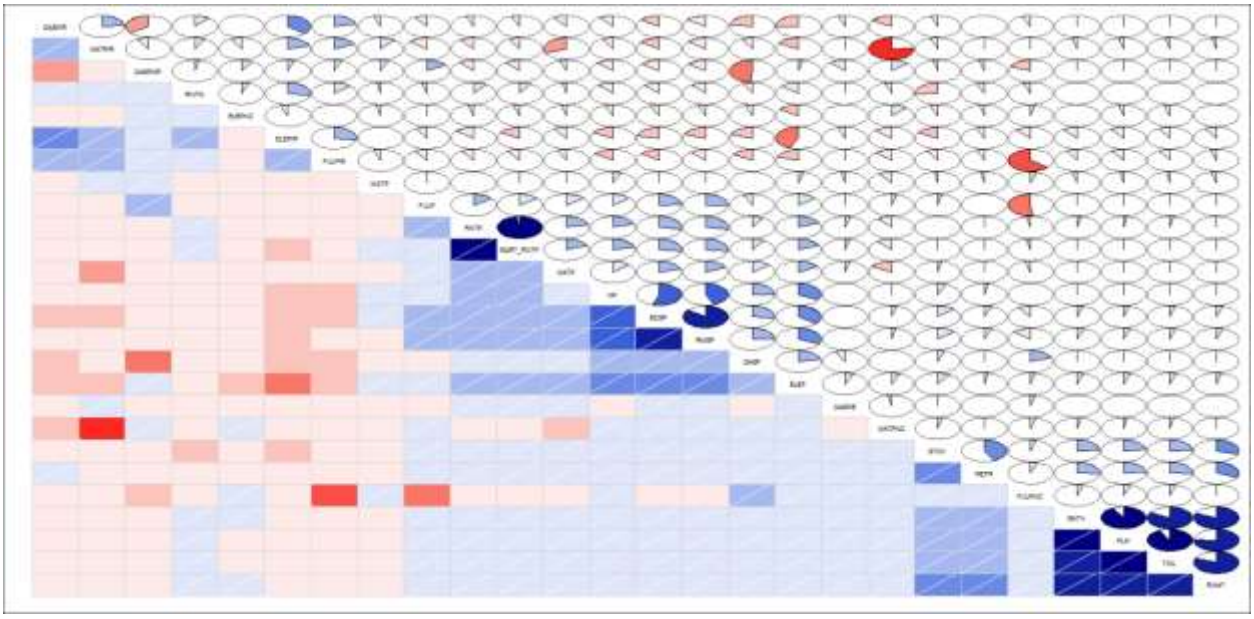
- ✓ From the above box plot we can see that there are many extreme observations in Units in Structure (BLD), Heating Fuel Type (HFL) and Lot Size (ACR) columns. This might affect linear regression results if we include these variables into the model.
- ✓ There are very less households with heating fuel type 5 and 7.
- ✓ Maximum variation in rent is observed in units in structure column (BLD).

Visualization #2: Continuous variables with Rent Amount - Scatter Plot



- ✓ From the above scatter plot we can see that rent does not really depend on number of rooms, and in the dataset it is not mentioned that the full housing property is rented or few rooms are rented. So e cannot derive an inference that number of rooms are not affecting the rent at all. Same goes with number of people living in the apartment.

Visualization #3: Continuous variables with Rent Amount - Corrgram (Correlation Analysis)



From the above Corrgram we can analyze what all variables are correlated:

For example, graph shows that PLM is highly correlated to BATH. From the data description of these column, the high correlation between them makes sense.

Variable Name	Levels	Description
BATH	N/A	Group Quarters

(Bathtub or Shower)	1	Yes
	2	No
PLM (hot and cold running water, a flush toilet, and a bathtub or shower)	N/A	Group Quarters
	1	Yes
	2	No

Splitting data into train and test dataset

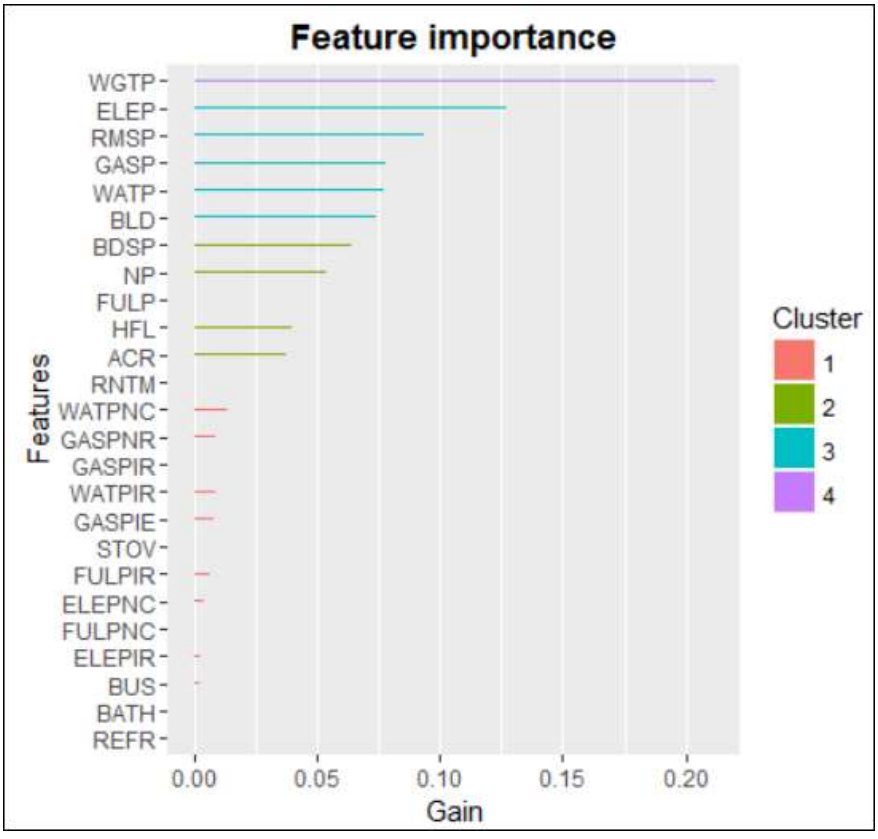
- ✓ I have partitioned the data into 2 datasets, training and test dataset
- ✓ Train dataset consists of 75% of the data and test dataset consists of 25% of the data

Non-Parametric Modeling Using Extreme Gradient Boosting

I have used extreme gradient boosting using 'xgboost' package in R and results are as follows:

```
> NumMetrics(testRentalData$RNTP,testRentalData$predXgboost)
          MAD           MSE          MAPE          MPSE          tMAD           p90            R2
3.628563e+02 2.492895e+05 8.571226e+01 5.255132e+03 3.251903e+02 8.051166e+02 1.777883e-01
```

As we can see that the model explains approximately 18% (R2 is 0.177) of the variance in rent amount. I have identified influential variable which could be used in the regression model as follows:

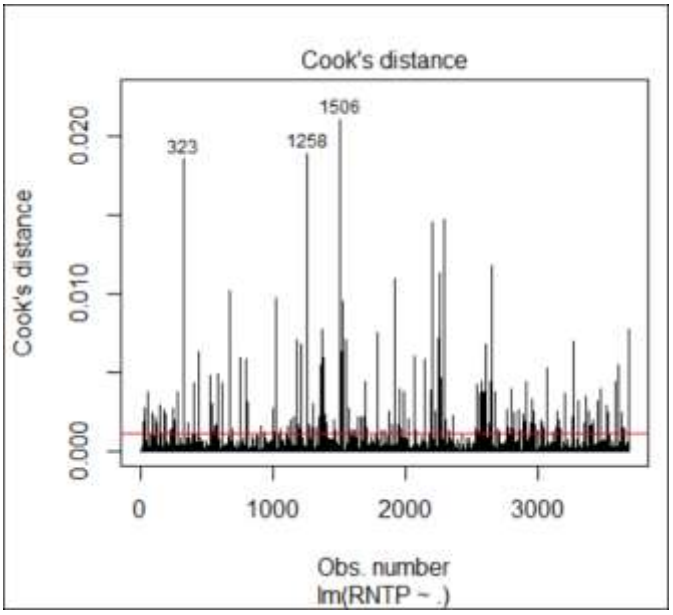


I have calculated top 10 features based on gain as follows:

```
> importance_matrix[order(importance_matrix[,c(1:10)]),]$Feature
 [1] "WGTP" "ELEP" "RMSP" "GASP" "WATP" "BLD"  "BDSP" "NP"   "FULP" "HFL"
```

Parametric Modeling Using Linear Regression

Extreme Value Analysis (Influential Observation Analysis using Cooks Distance)



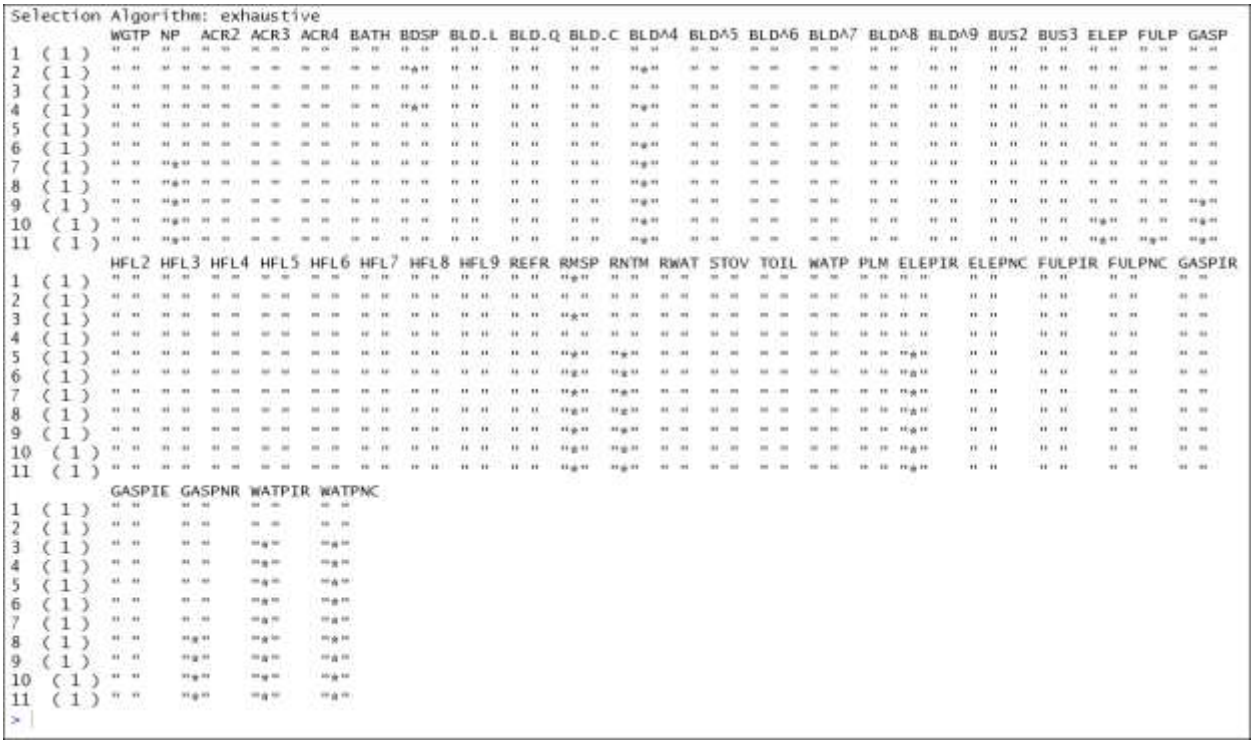
I have calculated cooks distance to identify extreme values that affects the parameter estimates of linear regression model

I have calculated total percentage of influential observations and found that total 5.54% of the observations are influential observations.

Best Subsets Regression Modeling

I have used regsubsets function from the leaps package to identify the key features that affects the monthly rent as follows:

```
#########################################################################################################
# Best Predictor variables based on regsubsets function from leaps package
# (Model selection by exhaustive search, forward or backward stepwise, or sequential replacement)
#########################################################################################################


regfit = regsubsets(SQRT_RNTP~.-RNTP,data=CTHousingRentalFeatures,nvmax = 10)
summary(regfit)
```

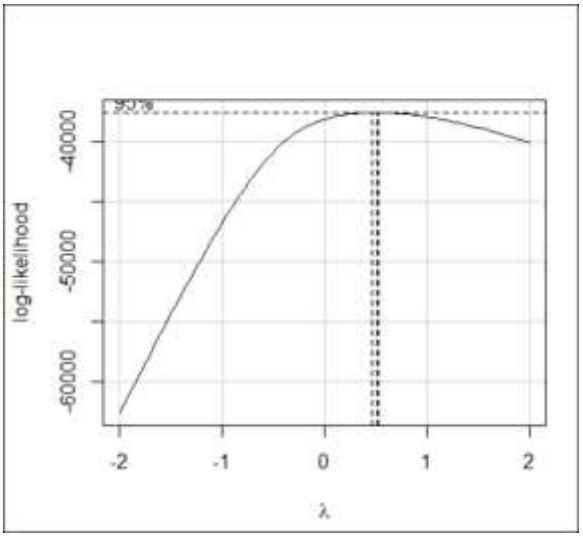


From the summary, we can see that following variables are best predictor variables RMSP, WATPIR, WATPNC, BLD^4, RNTM, ELEPIR, NP, GASPNR

Variable Transformation

I have used box cox method of MASS package to identify what would be the best transformation applied to target variable (Monthly Rent) so that the relationship between target variable and predictors become linear.

```
# Finding potential transformation for Target Variable RNTP
reg1 = lm(RNTP~.,data = CTHousingRentalFeatures)
boxCox(reg1,family="yjPower",plotit = T)
```



- ✓ From the plot, it is evident that we can apply squared root transformation to our Y variable to make relationship linear
- ✓ I have created new variable with squared root of RNTP (Monthly Rent)

Regression Models

Based on the best subsets regression modeling results I have used those variables to create linear regression model as follows:

```
#############################################################
linearReg <- glm(SQRT_RNTP~RMSP+WATPIR+WATPNC+BLD+RNTM+ELEPIR+NP+GASPNR, data=trainRentalData)
```

From the extreme value analysis, we found that the dataset consists of 5.54% of the observations with influential values. In this case, we should use robust regression model which assigns allocates weights to the observations based on Cooks Distance and gives less weightage to the influential observations.

I have created robust linear regression model as follows:

```
robustReg <- rlm(SQRT_RNTP~RMSP+WATPIR+WATPNC+RNTM+ELEPIR+NP+GASPNR, data=trainRentalData)
```

Two-Way Interactions between Predictors

I tried to identify if there is any significant interaction between any two variables which could be included to the model to increase the model performance, however from the results I found that there are no such interactions that can enhance model performance.

```
#################################################################
# Identifying 2 way interactions for regression model
#################################################################

# Checks all 2 way interactions
res = step(linearRegUpdated,~.^2)

res$anova
# There are no such significant interactions which we can add in our model to improve accuracy
```

```
> res$anova
               Step Df Deviance Resid. Df Resid. Dev      AIC
1                   NA       NA      2752   180988.7 19396.18
2      + WATPNC:NP -1 902.8108      2751   180085.8 19384.38
3    + RNTM:ELEPIR -1 489.3578      2750   179596.5 19378.87
4      + ELEPIR:NP -1 504.1875      2749   179092.3 19373.11
5  + WATPNC:GASPNR -1 269.3901      2748   178822.9 19370.96
6    + RMSP:WATPNC -1 299.1134      2747   178523.8 19368.34
7  + WATPIR:GASPNR -1 263.8984      2746   178259.9 19366.25
8      + NP:GASPNR -1 307.1065      2745   177952.8 19363.49
9      + RMSP:RNTM -1 251.9876      2744   177700.8 19361.58
10   + RMSP:WATPIR -1 272.9391      2743   177427.9 19359.34
11       + RNTM:NP -1 255.6672      2742   177172.2 19357.36
12 + WATPIR:ELEPIR -1 128.4876      2741   177043.7 19357.36
13 + WATPNC:ELEPIR -1 201.3890      2740   176842.3 19356.22
14   + RMSP:ELEPIR -1 157.3472      2739   176685.0 19355.76
15 + ELEPIR:GASPNR -1 138.2906      2738   176546.7 19355.60
16   + RNTM:GASPNR -1 148.8708      2737   176397.8 19355.27
```

Regression Model Performance Evaluation

After creating two regression models I have evaluated the model performance using the function that I have developed as follows:

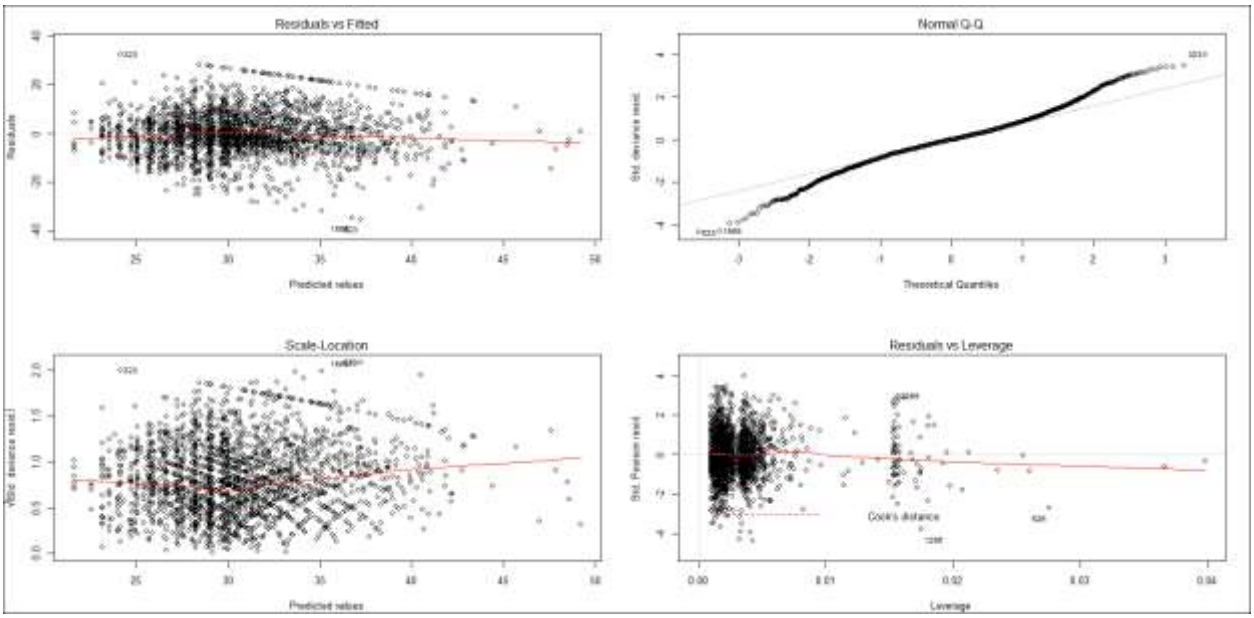
```
NumMetrics = function(a,m)  #here a is target value matrix and m is model value matrix
{
  metrics = c(MAD=0, MSE = 0, MAPE = 0, MPSE = 0, tMAD = 0, p90=0, R2 =0)
  metrics["MAD"] = mean(abs(a-m))
  metrics["MSE"] = mean((a-m)^2)
  metrics["MAPE"] = mean(abs(a-m)*100/a)
  metrics["MPSE"] = mean((((a-m)/a)^2)*100)
  metrics["tMAD"] = mean(abs(a-m),trim = 0.05)
  metrics["p90"] = quantile(abs(a-m),probs = 0.9)

  SST = sum((a-mean(a))^2)
  SSE = sum((a-m)^2)
  metrics["R2"] = 1- (SSE/SST)
  return(metrics)
}
```

```
> NumMetrics(testRentalData$RNTP,testRentalData$robustRegPred)
         MAD          MSE         MAPE         MPSE         tMAD          p90           R2
3.565956e+02 2.592435e+05 8.092276e+01 4.603913e+03 3.128036e+02 8.009205e+02 1.449578e-01
> NumMetrics(testRentalData$RNTP,testRentalData$linearRegPred)
         MAD          MSE         MAPE         MPSE         tMAD          p90           R2
3.574385e+02 2.596785e+05 8.106851e+01 4.483672e+03 3.138774e+02 7.999893e+02 1.435229e-01
```

From the above evaluation, it is evident that using robust regression actually not helping us in this case, because the model with robust linear regression is showing almost identical results with linear regression.

From the model evaluation, we can conclude that using either of the model will not increase the model accuracy. Thus, we can use linear regression model for our further analysis.

Linear Regression Model Diagnostics



The normal Q-Q plot suggests that the model residuals are not normally distributed.

I have performed Shapiro Test to confirm it:

```
> shapiro.test(x = linearRegUpdated$residuals)

        Shapiro-Wilk normality test

data:  linearRegUpdated$residuals
W = 0.98064, p-value < 2.2e-16
```

- ✓ p value is less than 0.05 thus we can reject the null hypothesis that model residuals are normally distributed

I have calculated variance inflation factor as follows:

```
> vif(linearRegUpdated)
    RMSP    WATPIR    WATPNC      RNTM    ELEPIR        NP    GASPNR
1.305260 2.455664 2.384766 1.093583 1.161290 1.267235 1.047361
```

- ✓ The results show that (values less than 4) there is no multicollinearity present in the data.

Linear Regression Model Result Interpretation

```
> summary(linearRegUpdated)

Call:
glm(formula = SQRT_RNTP ~ RMSP + WATPIR + WATPNC + RNTM + ELEPIR +
    NP + GASPNR, data = trainRentalData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-35.214   -4.458    0.000    4.266   32.428

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.69791    0.64925  47.282  < 2e-16 ***
RMSP         0.94930    0.09439  10.057  < 2e-16 ***
WATPIR      -5.16940    0.49269 -10.492  < 2e-16 ***
WATPNC      -6.71848    0.53045 -12.666  < 2e-16 ***
RNTM         9.56866    1.04133   9.189  < 2e-16 ***
ELEPIR      -3.11744    0.48547  -6.421 1.58e-10 ***
NP           0.64633    0.13180   4.904 9.95e-07 ***
GASPNR      -0.81529    0.31805  -2.563   0.0104 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 65.76623)

    Null deviance: 219101  on 2759  degrees of freedom
Residual deviance: 180989  on 2752  degrees of freedom
AIC: 19396

Number of Fisher Scoring iterations: 2
```

Since in this model we are predicting squared root of the monthly rent we need to square the estimates to interpret their effect on the monthly rent.

```
> round(coef(linearRegUpdated),2)^2
(Intercept)        RMSP      WATPIR      WATPNC        RNTM      ELEPIR          NP      GASPNR
   942.4900      0.9025     26.7289     45.1584     91.5849      9.7344      0.4225      0.6724
```

- ✓ The estimated rate of change of the squared root of conditional mean of Monthly Rent with respect to Number of Rooms, when all other predictors are fixed is 0.9025$.
- ✓ As we can see that it is hard to interpret this model because of the squared root transformation, in this case we can use normal linear regression model for the parameter estimates

```
> linearReg <- glm(RNTP~RMSP+WATPIR+WATPNC+BLD+RNTM+ELEPIR+NP+GASPNR, data=trainRentalData)
> summary(linearReg)

Call:
glm(formula = RNTP ~ RMSP + WATPIR + WATPNC + BLD + RNTM + ELEPIR +
    NP + GASPNR, data = trainRentalData)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1536.99   -313.13    -60.97    213.74   2551.45

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  930.769     56.504  16.473  < 2e-16 ***
RMSP          72.811      6.657  10.938  < 2e-16 ***
WATPIR      -383.994     32.053 -11.980  < 2e-16 ***
WATPNC      -466.816     34.367 -13.583  < 2e-16 ***
BLD           13.929      5.279   2.639  0.00837 **
RNTM         629.229     67.621   9.305  < 2e-16 ***
ELEPIR      -162.132     31.456  -5.154 2.73e-07 ***
NP            41.211      8.564   4.812 1.57e-06 ***
GASPNR       -48.292     20.506  -2.355  0.01859 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 273245)

    Null deviance: 919606227  on 2759  degrees of freedom
Residual deviance: 751697041  on 2751  degrees of freedom
AIC: 42394

Number of Fisher Scoring iterations: 2
```

```
> NumMetrics(trainRentalData$RNTP,  predict(linearReg,testRentalData))
         MAD          MSE         MAPE         MPSE         tMAD          p90           R2
4.487670e+02 3.831254e+05 1.309971e+02 8.130258e+03 3.999731e+02 9.503831e+02 -1.498684e-01
```

Model Parameter Estimates

- ✓ A unit increase in number of rooms, keeping everything else constant will translate into 72.81$ increase in the rent of the housing property
- ✓ Including water yearly cost in rent or condo fees, keeping everything else constant will translate into 383.99$ decrease in the rent of housing property
- ✓ If there are no water consumption charges, keeping everything else constant will translate into 466.82$ decrease in the rent of housing property
- ✓ A unit increase in unit structure, keeping everything else constant will translate into 13$ increase in the rent of housing property
- ✓ If meals are included in the rent, keeping everything else constant will translate into 629.22$ increase in the rent of housing property
- ✓ Including electricity monthly cost in rent or condo fees, keeping everything else constant will translate into 162.13$ decrease in the rent of housing property
- ✓ If number of person living in the housing property increases by one, keeping everything else constant will translate into 41.21$ increase in the rent of housing property
- ✓ If there are no gas consumption monthly charges, keeping everything else constant will translate into 48.29$ decrease in the rent of housing property

Conclusion

In this assignment, I have used both parametric and non-parametric modeling techniques to achieve maximum information gain from the given dataset, however it seems that we need more variables to predict monthly rent accurately.

Although the R2 value is about 0.14-0.20, the linear regression model serves the purpose of exploratory analysis and provides us the factors that influences the monthly rent.

Appendix

Data Dictionary

Variable Name	Levels	Description
SERIALNO - Housing unit/GQ person serial number	0000001..9999999	Housing unit/GQ person serial number (Unique identifier)
DIVISION - Division code	0	Puerto Rico
	1	New England (Northeast region)
	2	Middle Atlantic (Northeast region)
	3	East North Central (Midwest region)
	4	West North Central (Midwest region)
	5	South Atlantic (South region)
	6	East South Central (South region)
	7	West South Central (South Region)
	8	Mountain (West region)
	9	Pacific (West region)
Region	1	Northeast
	2	Midwest
	3	South
	4	West
	9	Puerto Rico
WGTP - Housing Weight	00000	Group Quarters place holder record
	00001..09999	Integer weight of housing unit
NP -Number of person records following this housing record	1	Vacant unit
	01	One person record (one person in household or any person in group quarters)
	02..20	Number of person records (number of persons in household)
TYPE - Type of unit	1	Housing unit
	2	Institutional group quarters
	3	Non-institutional group quarters

ACR - Lot size	b	N/A (GQ/not a one-family house or mobile home)
	1	House on less than one acre
	2	House on one to less than ten acres
	3	House on ten or more acres
BATH - Bathtub or shower	b	N/A (GQ)
	1	Yes
	2	No
BDSP - Number of bedrooms	bb	N/A (GQ)
	00..99	0 to 99 bedrooms (Top-coded)
BUS - Business or medical office on property	b	N/A (GQ/not a one-family house or mobile home)
	1	Yes
	2	No
CONP - Condo fee (monthly amount)	b	N/A (GQ/vacant units, except “for sale-only” and “sold, not occupied”/not owned or being bought)
	0000	Not condo (Owned/being bought)
	0001..9999	0001..9999 .$1 - $9999 (Rounded and top-coded)
ELEP - Electricity (monthly cost)	B	N/A (GQ/vacant)
	1	Included in rent or in condo fee
	2	No charge or electricity not used
	3	003..999 .$3 to $999 (Rounded and top-coded)
FULP - Fuel cost (yearly cost for fuels other than gas and electricity)	b	N/A (GQ/vacant)
	1	Included in rent or in condo fee
	2	No charge or these fuels not used
	3	0003..9999 .$3 to $9999 (Rounded and top-coded)
GASP - Gas (monthly cost)	bbb	bbb .N/A (GQ/vacant)

	1	001 .Included in rent or in condo fee
	2	Included in electricity payment
	3	No charge or gas not used
	004..999	004..999 .$4 to $999 (Rounded and top-coded)
HFL - House heating fuel	b	N/A (GQ/vacant)
	1	Utility gas
	2	Bottled, tank, or LP gas
	3	Electricity
	4	Fuel oil, kerosene, etc
	5	Coal or coke
	6	Wood
	7	Solar energy
	8	Other fuel
	9	No fuel used
INSP - Fire/hazard/flood insurance (yearly amount)	bbbbb	N/A (GQ/vacant/not owned or being bought)
	00000	None
	00001-10000	$1 to $10000 (Rounded and top-coded)
REFR - Refrigerator	b	N/A (GQ)
	1	Yes
	2	No
RMSP - Number of Rooms	bb	N/A (GQ)
	00-99	Rooms (Top-coded)
RNTM Meals included in rent	b	N/A (GQ/vacant units, except “for rent” and “rented,not occupied”/owned or being bought /occupied without rent payment)
	1	Yes
	2	No
RNTP - Monthly rent	bbbbb	N/A (GQ/vacant units, except “for rent” and “rented, not

		occupied”/owned or being bought /occupied without rent payment)
	00001-99999	$1 to $99999 (Rounded and top-coded)
RWAT - Hot and cold running water	b	N/A (GQ)
	1	Yes
	2	No
	9	Case is from Puerto Rico, RWAT not applicable
STOV - Stove or range	b	N/A (GQ)
	1	Yes
	2	No
TEN - Tenure	b	N/A (GQ/vacant)
	1	Owned with mortgage or loan (include home equity loans)
	2	Owned free and clear
	3	Rented
	4	Occupied without payment of rent
TOIL - Flush toilet	b	N/A (GQ)
	1	Yes
	2	No
VACS - Vacancy status	b	N/A (GQ/occupied)
	1	For rent
	2	Rented, not occupied
	3	For sale only
	4	Sold, not occupied
	5	For seasonal/recreational/occasional use
	6	For migrant workers
	7	Other vacant
VALP - Property value	NA	GQ/vacant units, except” for-sale-only“ and ”sold, not occupied”/not owned or being
	$1 to $9999999	(Rounded and top-coded)

WATP - Water (yearly cost)	bbbb	N/A (GQ/vacant)
	0001	Included in rent or in condo fee
	0002	No charge
	0003-9999	$3 to $9999 (Rounded and top-coded)
HOTWAT - Water heater (Puerto Rico only)	b	N/A (GQ)
	1	Yes
	2	No
	9	Case is from the United States, HOTWAT not applicable
KIT - Complete kitchen facilities	b	N/A (GQ)
	1	Yes, has stove or range, refrigerator, and sink with a faucet
	2	No
PLM - Complete plumbing facilities	b	N/A (GQ)
	1	Yes, has hot and cold running water, a flush toilet, and a bathtub or shower
	2	No
	9	Case is from Puerto Rico, PLM recode not applicable
PLMPRP - Complete plumbing facilities for Puerto Rico	b	N/A (GQ)
	1	Yes, has running water, a flush toilet, and a bathtub or shower
	2	No
	9	Case is from the United States, PLMPRP not applicable
SRNT - Specified rent unit	b	N/A
	0	Not specified rent unit
	1	Specified rent unit
MIG - Mobility status (lived here 1 year ago)	b	N/A (less than 1 year old)
	1	Yes, same house (nonmovers)
	2	No, outside US and Puerto Rico

	3	No, different house in US or Puerto Rico