MSA-2020

Helitha Dharmadasa

26/07/2020

Importing and Cleaning

```
houses.df = read.table("Dataset_Final.csv", header=T, sep=",")
head(houses.df)

##   Bedrooms Bathrooms                                    Address Land.area
## 1        5         3  106 Lawrence Crescent Hill Park, Auckland       714
## 2        5         3             8 Corsica Way Karaka, Auckland       564
## 3        6         4     243 Harbourside Drive Karaka, Auckland       626
## 4        2         1  2/30 Hardington Street Onehunga, Auckland        65
## 5        3         1     59 Israel Avenue Clover Park, Auckland       601
## 6        3         1 14 Tainui Terrace Mangere Bridge, Auckland       100
##        CV  Latitude Longitude     SA1 X0.19.years X20.29.years
X30.39.years
## 1  960000 -37.01292  174.9041 7009770          48           27
24
## 2 1250000 -37.06367  174.9229 7009991          42           18
12
## 3 1250000 -37.06358  174.9240 7009991          42           18
12
## 4  740000 -36.91300  174.7874 7007871          42            6
21
## 5  630000 -36.97904  174.8926 7008902          93           27
33
## 6 1050000 -36.94393  174.7805 7007917          63           15
24
##   X40.49.years X50.59.years X60..years        Suburbs Population
## 1           21           24         21       Manurewa        174
## 2           21           15         30         Karaka        129
## 3           21           15         30         Karaka        129
## 4           21           12         15       Onehunga        120
## 5           30           21         33    Clover Park        231
## 6           33           30         39 Mangere Bridge        195
##   Deprivation.Index
## 1                 6
## 2                 1
## 3                 1
## 4                 2
## 5                 9
## 6                 4

summary(houses.df)
```

```
##     Bedrooms        Bathrooms       Address           Land.area
##  Min.   : 1.000   Min.   :1.000   Length:1051        Length:1051
##  1st Qu.: 3.000   1st Qu.:1.000   Class :character   Class :character
##  Median : 4.000   Median :2.000   Mode  :character   Mode  :character
##  Mean   : 3.777   Mean   :2.073
##  3rd Qu.: 4.000   3rd Qu.:3.000
##  Max.   :17.000   Max.   :8.000
##                   NA's   :2
##         CV              Latitude        Longitude          SA1
##  Min.   :  270000   Min.   :-37.27   Min.   :174.3   Min.   :7001130
##  1st Qu.:  780000   1st Qu.:-36.95   1st Qu.:174.7   1st Qu.:7004416
##  Median : 1080000   Median :-36.89   Median :174.8   Median :7006325
##  Mean   : 1387521   Mean   :-36.89   Mean   :174.8   Mean   :7006319
##  3rd Qu.: 1600000   3rd Qu.:-36.86   3rd Qu.:174.9   3rd Qu.:7008384
##  Max.   :18000000   Max.   :-36.18   Max.   :175.5   Max.   :7011028
##
##   X0.19.years      X20.29.years     X30.39.years     X40.49.years
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
##  1st Qu.: 33.00   1st Qu.: 15.00   1st Qu.: 15.00   1st Qu.: 18.00
##  Median : 45.00   Median : 24.00   Median : 24.00   Median : 24.00
##  Mean   : 47.55   Mean   : 28.96   Mean   : 27.04   Mean   : 24.13
##  3rd Qu.: 57.00   3rd Qu.: 36.00   3rd Qu.: 33.00   3rd Qu.: 30.00
##  Max.   :201.00   Max.   :270.00   Max.   :177.00   Max.   :114.00
##
##   X50.59.years     X60..years       Suburbs            Population
##  Min.   : 0.00   Min.   :  0.00   Length:1051        Min.   :  3.0
##  1st Qu.:15.00   1st Qu.: 18.00   Class :character   1st Qu.:138.0
##  Median :21.00   Median : 27.00   Mode  :character   Median :174.0
##  Mean   :22.62   Mean   : 29.36                      Mean   :179.9
##  3rd Qu.:27.00   3rd Qu.: 36.00                      3rd Qu.:210.0
##  Max.   :90.00   Max.   :483.00                      Max.   :789.0
##
##  Deprivation.Index
##  Min.   : 1.000
##  1st Qu.: 2.000
##  Median : 5.000
##  Mean   : 5.064
##  3rd Qu.: 8.000
##  Max.   :10.000
##

houses.df$Land.area = gsub("[^0-9]", "", houses.df$Land.area)
houses.df$Land.area = as.numeric(houses.df$Land.area)

for(i in 1:ncol(houses.df)){
  houses.df[is.na(houses.df[,i]), i] <- median(houses.df[,i], na.rm = TRUE)
}
```

We can see that the Land Area column is recorded as a string/char so we convert it to a numeric value by removing any non-numeric elements and converting with as.numeric().

We can see from the summary that there are 2 ‘NA’ values in the bathrooms column, so we execute some code to replace any ’NA’s found with the column median. As we run this after converting the Land Area column it will be checked as well.

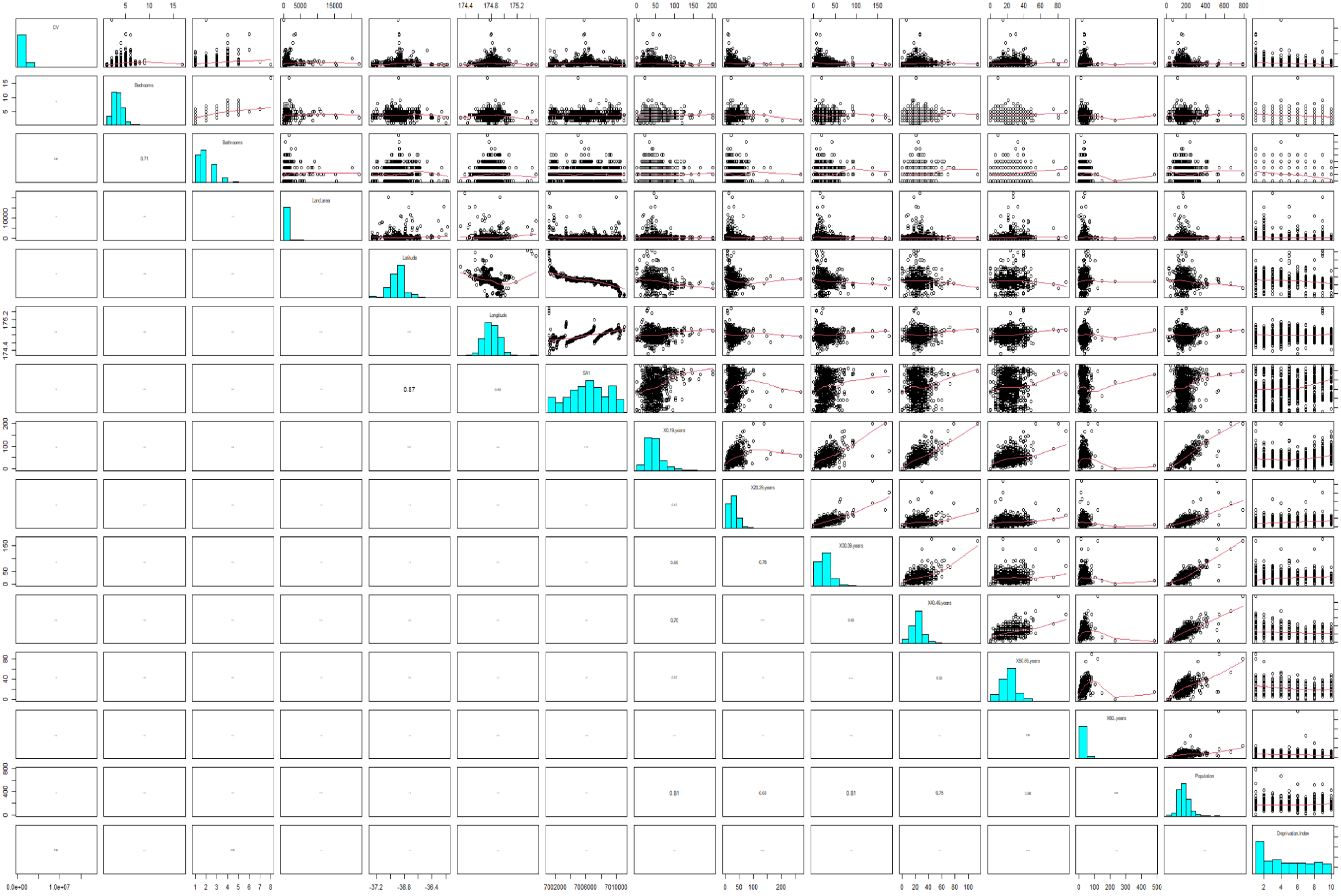
Initial Analysis and Thoughts

```
summary(houses.df)

##     Bedrooms        Bathrooms       Address            Land.area
##  Min.   : 1.000   Min.   :1.000   Length:1051        Min.   :   40
##  1st Qu.: 3.000   1st Qu.:1.000   Class :character   1st Qu.:  321
##  Median : 4.000   Median :2.000   Mode  :character   Median :  571
##  Mean   : 3.777   Mean   :2.073                      Mean   :  857
##  3rd Qu.: 4.000   3rd Qu.:3.000                      3rd Qu.:  825
##  Max.   :17.000   Max.   :8.000                      Max.   :22240
##        CV              Latitude        Longitude          SA1
##  Min.   :  270000   Min.   :-37.27   Min.   :174.3   Min.   :7001130
##  1st Qu.:  780000   1st Qu.:-36.95   1st Qu.:174.7   1st Qu.:7004416
##  Median : 1080000   Median :-36.89   Median :174.8   Median :7006325
##  Mean   : 1387521   Mean   :-36.89   Mean   :174.8   Mean   :7006319
##  3rd Qu.: 1600000   3rd Qu.:-36.86   3rd Qu.:174.9   3rd Qu.:7008384
##  Max.   :18000000   Max.   :-36.18   Max.   :175.5   Max.   :7011028
##   X0.19.years      X20.29.years     X30.39.years     X40.49.years
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
##  1st Qu.: 33.00   1st Qu.: 15.00   1st Qu.: 15.00   1st Qu.: 18.00
##  Median : 45.00   Median : 24.00   Median : 24.00   Median : 24.00
##  Mean   : 47.55   Mean   : 28.96   Mean   : 27.04   Mean   : 24.13
##  3rd Qu.: 57.00   3rd Qu.: 36.00   3rd Qu.: 33.00   3rd Qu.: 30.00
##  Max.   :201.00   Max.   :270.00   Max.   :177.00   Max.   :114.00
##   X50.59.years     X60..years       Suburbs            Population
##  Min.   : 0.00   Min.   :  0.00   Length:1051        Min.   :  3.0
##  1st Qu.:15.00   1st Qu.: 18.00   Class :character   1st Qu.:138.0
##  Median :21.00   Median : 27.00   Mode  :character   Median :174.0
##  Mean   :22.62   Mean   : 29.36                      Mean   :179.9
##  3rd Qu.:27.00   3rd Qu.: 36.00                      3rd Qu.:210.0
##  Max.   :90.00   Max.   :483.00                      Max.   :789.0
##  Deprivation.Index
##  Min.   : 1.000
##  1st Qu.: 2.000
##  Median : 5.000
##  Mean   : 5.064
##  3rd Qu.: 8.000
##  Max.   :10.000
```

All “NA’s” are gone, and all columns that need to be numeric are now numeric. From a glance we can see that there are quite large max values and potential outliers in the bedrooms, year columns and in the Population column and especially the CV column. The rest of the data seems more naturally spread with maximums close to their 3rd quartiles and means and medians. We will have to keep an eye out on the variables with these potential outliers when looking at the cooks plot when creating our linear model.

We create a pairs plot and exclude any non-numeric variables. These variables are excluded for the remaineder of the process and they cannot be handled through these processes

```
pairs20x(houses.df[c(5,1,2,4,6,7,8,9,10,11,12,13,14,16,17)])
```

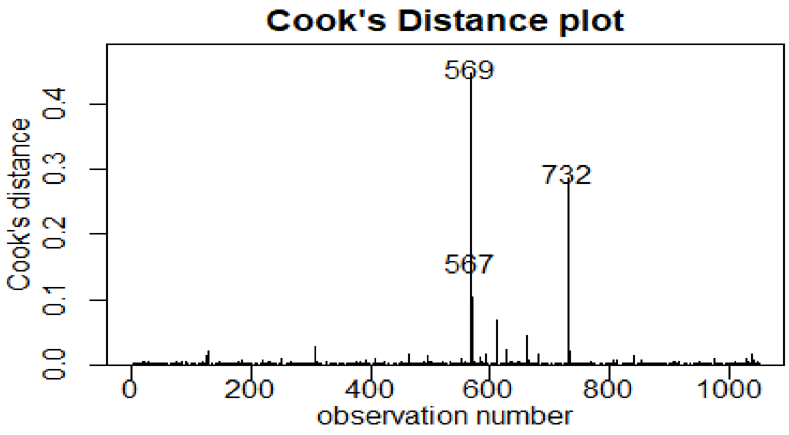

CV
Bedrooms
Bathrooms
Land.area
Latitude
Longitude
SA1
X0.19.years
X20.29.years
X30.39.years
X40.49.years
X50.59.years
X60..years
Population
Deprivation.Index
0.71
0.87
0.78
0.60
0.70
0.81
0.81
0.75
0.0e+00
1.0e+07
1.5e+07
1 2 3 4 5 6 7 8
-37.2
-36.8
-36.4
-36.6
174.4
174.8
175.2
7002000
7006000
7010000
0 5000 15000
0 10000
0 50 100 150 200
0 50 100 200
0 100 250
0 50 100 150
0 20 40 60 80 100
0 20 40 60 80
0 100 200 300 400 500
0 200 400 600 800
0 400 800
2 4 6 8 10
5 10 15

We can try build a linear model that uses the other variables to estimate the capital value of the property. We can do some obvious transformations like logging the price as they are usually Value doesn't seem to be too strongly correlated with anything this is supported by the distribution of CV above, population appears to have a strong correlation with quite a few other variables, Latitude and SA1 appear to have quite a strong correlation. CV doesn't appear to have a strong positive relationship with many other variables, at least at a glance. We can also include some guesses at potential interaction effects such as with bathrooms and bedrooms, and lat, long and land area. This is effectively going to be our worst case model to base the next section on.

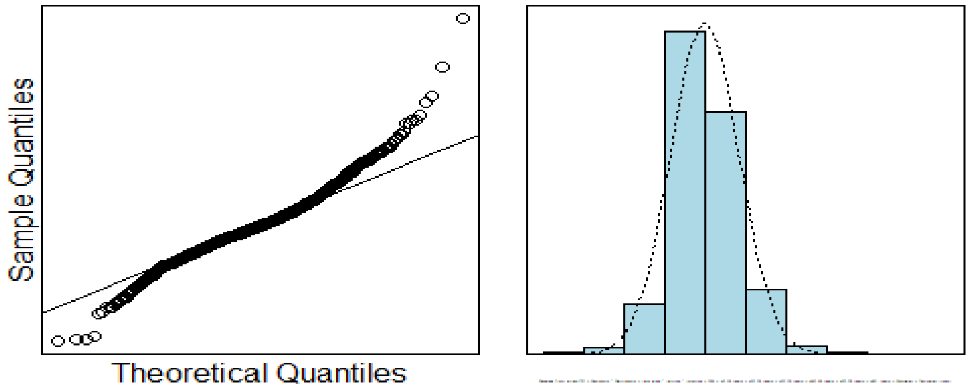
Fitting our Linear Model

```
houses.fit <- lm(log(CV) ~ Bedrooms *  Bathrooms + Land.area * Latitude *
Longitude + SA1 + X0.19.years + X20.29.years + X30.39.years + X40.49.years +
X50.59.years + X60..years + Population + Deprivation.Index, data=houses.df)

#Methods and Assumptions checks
summary(houses.fit)

##
## Call:
## lm(formula = log(CV) ~ Bedrooms * Bathrooms + Land.area * Latitude *
##     Longitude + SA1 + X0.19.years + X20.29.years + X30.39.years +
##     X40.49.years + X50.59.years + X60..years + Population +
Deprivation.Index,
##     data = houses.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5305 -0.2335 -0.0358  0.2074  2.6449
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -3.221e+04  5.229e+03  -6.159 1.05e-09 ***
## Bedrooms                       1.001e-01  2.338e-02   4.280 2.04e-05 ***
## Bathrooms                      2.430e-01  2.960e-02   8.207 6.72e-16 ***
## Land.area                      4.271e+00  1.069e+00   3.994 6.96e-05 ***
## Latitude                      -8.754e+02  1.418e+02  -6.172 9.71e-10 ***
## Longitude                      1.846e+02  2.989e+01   6.176 9.46e-10 ***
## SA1                           -5.558e-06  1.292e-05  -0.430  0.66703
## X0.19.years                    1.503e-03  2.712e-03   0.554  0.57951
## X20.29.years                   3.675e-03  2.742e-03   1.340  0.18038
## X30.39.years                  -2.835e-03  2.882e-03  -0.984  0.32556
## X40.49.years                   7.215e-04  3.318e-03   0.217  0.82788
## X50.59.years                   7.914e-03  2.978e-03   2.658  0.00799 **
## X60..years                     3.185e-03  2.605e-03   1.222  0.22181
## Population                    -2.533e-03  2.573e-03  -0.984  0.32526
## Deprivation.Index             -6.890e-02  6.065e-03 -11.360  < 2e-16 ***
## Bedrooms:Bathrooms            -2.369e-02  5.456e-03  -4.341 1.56e-05 ***
## Land.area:Latitude             1.159e-01  2.914e-02   3.978 7.43e-05 ***
## Land.area:Longitude           -2.447e-02  6.117e-03  -4.001 6.77e-05 ***
## Latitude:Longitude             5.009e+00  8.114e-01   6.174 9.59e-10 ***
## Land.area:Latitude:Longitude -6.642e-04  1.667e-04  -3.985 7.22e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4249 on 1031 degrees of freedom
## Multiple R-squared:  0.4477, Adjusted R-squared:  0.4375
## F-statistic: 43.99 on 19 and 1031 DF,  p-value: < 2.2e-16
```

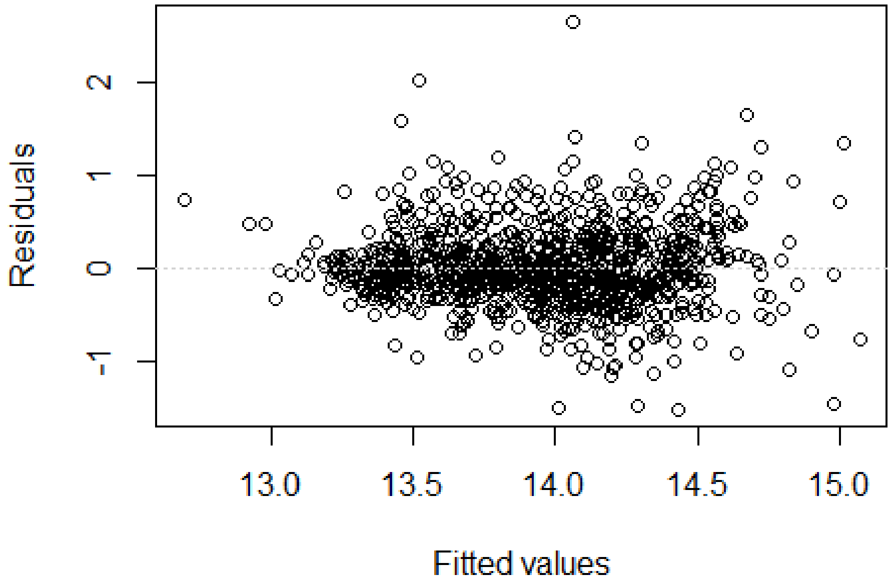
```
cooks20x(houses.fit)
```



```
normcheck(houses.fit)
```



```
eovcheck(houses.fit)
```



From this model we can see two things, one we have a lot of terms with poor significance and hence correlation in our model, and two there are some interaction effects present between some of or variables, specifically bathroom and bedrooms, and longitude and latitude and land area. We also have a few significant terms here we want to keep in the model. To simply the model down and hopefully improve its current poor R-squared value we can employ the MuMIn package. For the purposes of our methods and assumptions check we can see that normality looks good enough, EoV looks good and while there appears to be outliers at values 569, 732 and 567 in the cooks plot, they are within limits to leave in the model.

Simplifying our Model

```
options(na.action = "na.fail")

all.fits <- dredge(houses.fit)

## Fixed term is "(Intercept)"

head(all.fits)
```

```
## Global model call: lm(formula = log(CV) ~ Bedrooms * Bathrooms + Land.area
* Latitude *
##     Longitude + SA1 + X0.19.years + X20.29.years + X30.39.years +
##     X40.49.years + X50.59.years + X60..years + Population +
Deprivation.Index,
##     data = houses.df)
## ---
## Model selection table
##         (Int)    Bth     Bdr  Dpr.Ind Lnd.are    Ltt   Lng        Ppl
X0.19.yrs
## 513344 -32390 0.2456 0.10030 -0.06616   4.340 -879.3 185.4            -
0.001650
## 521856 -32230 0.2430 0.10010 -0.06855   4.251 -875.1 184.5 -0.0013190
## 515136 -33760 0.2424 0.09457 -0.07134   4.466 -916.6 193.3
## 513856 -32290 0.2446 0.09942 -0.06779   4.304 -876.7 184.9            -
0.001551
## 515392 -32820 0.2436 0.09895 -0.06840   4.358 -891.1 187.9            -
0.001146
## 513408 -32210 0.2455 0.10090 -0.06609   4.305 -874.4 184.4  0.0004725 -
0.002173
##        X20.29.yrs X30.39.yrs X40.49.yrs X50.59.yrs X60..yrs  Bth:Bdr
## 513344            -0.004469             0.005961          -0.02401
## 521856   0.002475  -0.003998             0.006857  0.00195 -0.02374
## 515136            -0.004592  -0.003272   0.005681          -0.02298
## 513856   0.001161  -0.005519             0.005783          -0.02389
## 515392            -0.004252  -0.001900   0.006236          -0.02366
## 513408            -0.005431             0.005027          -0.02407
##        Lnd.are:Ltt Lnd.are:Lng Ltt:Lng Lnd.are:Ltt:Lng df   logLik   AICc
delta
## 513344      0.1178    -0.02486   5.032      -0.0006750 16 -583.660 1199.8
0.00
## 521856      0.1154    -0.02436   5.008      -0.0006612 18 -581.776 1200.2
0.37
## 515136      0.1212    -0.02559   5.245      -0.0006947 16 -583.946 1200.4
0.57
## 513856      0.1168    -0.02466   5.017      -0.0006695 17 -582.963 1200.5
0.67
## 515392      0.1183    -0.02497   5.099      -0.0006779 17 -583.214 1201.0
1.17
## 513408      0.1169    -0.02467   5.004      -0.0006697 17 -583.285 1201.2
1.32
##        weight
## 513344  0.229
## 521856  0.190
## 515136  0.172
## 513856  0.163
## 515392  0.127
## 513408  0.118
## Models ranked by AICc(x)
```

The dredge function from MuMIn effectively tests every possible combination of variables present in our initial model to brute-force find the best simplification of our initial model based on AICc values to compare models. All variables dropped are not significant went explaining CV.

```
first.model <- get.models(all.fits,1)[[1]]
summary(first.model)
```

```
## 
## Call:
## lm(formula = log(CV) ~ Bathrooms + Bedrooms + Deprivation.Index + 
##     Land.area + Latitude + Longitude + X0.19.years + X30.39.years + 
##     X50.59.years + Bathrooms:Bedrooms + Land.area:Latitude + 
##     Land.area:Longitude + Latitude:Longitude + 
Land.area:Latitude:Longitude + 
##     1, data = houses.df)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.54137 -0.23069 -0.04134  0.20596  2.66447 
## 
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                  -3.239e+04  5.191e+03  -6.239 6.40e-10 ***
## Bathrooms                     2.456e-01  2.950e-02   8.325 2.65e-16 ***
## Bedrooms                      1.003e-01  2.323e-02   4.317 1.73e-05 ***
## Deprivation.Index            -6.617e-02  5.437e-03 -12.170  < 2e-16 ***
## Land.area                     4.340e+00  1.067e+00   4.069 5.09e-05 ***
## Latitude                     -8.793e+02  1.409e+02  -6.242 6.31e-10 ***
## Longitude                     1.854e+02  2.969e+01   6.245 6.19e-10 ***
## X0.19.years                  -1.650e-03  7.883e-04  -2.093 0.036579 *  
## X30.39.years                 -4.470e-03  9.328e-04  -4.791 1.90e-06 ***
## X50.59.years                  5.961e-03  1.618e-03   3.685 0.000241 ***
## Bathrooms:Bedrooms           -2.401e-02  5.436e-03  -4.417 1.11e-05 ***
## Land.area:Latitude            1.178e-01  2.906e-02   4.053 5.43e-05 ***
## Land.area:Longitude          -2.486e-02  6.101e-03  -4.075 4.95e-05 ***
## Latitude:Longitude            5.032e+00  8.058e-01   6.244 6.20e-10 ***
## Land.area:Latitude:Longitude -6.750e-04  1.662e-04  -4.060 5.27e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.4247 on 1036 degrees of freedom
## Multiple R-squared:  0.4455, Adjusted R-squared:  0.438 
## F-statistic: 59.45 on 14 and 1036 DF,  p-value: < 2.2e-16
```

```
100*(exp(confint(first.model))-1)
```

```
##                                      2.5 %        97.5 %
## (Intercept)                  -1.000000e+02 -1.000000e+02
## Bathrooms                     2.064603e+01  3.545422e+01
## Bedrooms                      5.623080e+00   1.570602e+01
## Deprivation.Index            -7.395584e+00  -5.398472e+00
## Land.area                     8.455636e+02   6.207601e+04
## Latitude                     -1.000000e+02  -1.000000e+02
## Longitude                     1.661819e+57  6.720588e+107
## X0.19.years                  -3.191777e-01  -1.031753e-02
## X30.39.years                 -6.280171e-01  -2.635541e-01
## X50.59.years                  2.790669e-01   9.177971e-01
## Bathrooms:Bedrooms           -3.408385e+00  -1.325512e+00
## Land.area:Latitude            6.265881e+00   1.910493e+01
## Land.area:Longitude          -3.616604e+00  -1.280943e+00
## Latitude:Longitude            3.052338e+03   7.439198e+04
## Land.area:Latitude:Longitude -1.000694e-01  -3.486984e-02
```

Here we can see that in our best model produced by dredge all the terms included are now significant, and our R-squared improved marginally, though it will still be poor for prediction. This will be our final model.

Methods and Assumptions Checks

From previous experience with price variables as well as analysis of the pairs plot it was reasoned that the CV variable should be log transformed. Following this we fit a 'worst case model' without overdoing it (i.e. fitting interactions on all terms) as a baseline for our dredge, which we used to automatically simplify the model. Normality can be seen to be satisfactory but not perfect, EoV lacks any obvious curvature or trend, and outliers in the cooks plot remain within limits to leave in the model.

Executive Summary and Conclusions

My interest in this data was to see how the CV or capital value (i.e. price) of a property in NZ could be explained by the other variables in the data set.

Our model only 45% of the variance in the data and therefore is not suited for prediction.

The presence of interactions makes it difficult to calculate exact % increases, however generally the number of bathrooms and bedrooms increase value, strangely deprivation index being higher seems to decrease value, area seems to increase value, however the triple interaction with longitude and latitude here complicates matters, having people ages 0-19 decreases value by between 31.9 and 103%, ages 30-30 decreases value by 62.8 and 26.3% while ages 50-59 appear to increase value by 27 to 91%.