HW3

Lauren Stover

3/31/2021

ECO 395M: Exercises 3

Due date: Friday, April 9, 9 AM US Central Time

What causes what?

First, listen to this podcast from Planet Money. Then use your knowledge of statistical learning to answer the following questions.

1. Why can't I just get data from a few different cities and run the regression of "Crime" on "Police" to understand how more cops in the streets affect crime? ("Crime" refers to some measure of crime rate and "Police" measures the number of cops in a city.)
   -You can't just get the data from a few different cities and run crime on police because the number of police officers dispatched at any given time may be due to other reasons. It may be due to a terrorist threat in D.c., it may be due to a holiday in a major city so they dispatch more officers on any given day.

2. How were the researchers from UPenn able to isolate this effect? Briefly describe their approach and discuss their result in the "Table 2" below, from the researchers' paper.

-The researchers were able to isolate the effect of the high alert by looking at the midday ridership. By controlling for midday ridership, one could understand if due to the high alert days there was less crime because there were less civilians. As seen in Table 2, the civilian ridership was similar on days when high alert was issue and therefore ridership was significant at the 1 percent level while high alert was significant at the 5 percent level.

3. Why did they have to control for Metro ridership? What was that trying to capture?
   -The researchers controlled for metro ridership because they suspected on days where more police are dispatched within D.C. due to higher terrorist threats, there may be less civilians which could contribute to less victims for crimes. The Metro ridership indicated that civilians on these higher terrorist threat days were likely similar in levels due to the Metro ridership.

4. Below I am showing you "Table 4" from the researchers' paper. Just focus on the first column of the table. Can you describe the model being estimated here? What is the conclusion?

The model being estimated here is a linear model that includes the daily total number of crimes being predicted by high alert interacted with the dummy variable that is district 1, high alert being interacted with other distracts, a log of the midday ridership and the constant term that absorbs unobservable effects. Here we can see that the correlation between the daily rate of crimes and the high alert interacted with crime in the dummy variable is significant at the 1% level. This means that when there is a high alert in this

district, the total daily amount of crimes is extremely correlated. The log of midday ridership is correlated with daily crime at the 5% level, which makes sense. As the number of civilians increases, crime will be able to increase.

Predictive model building: green certification

Consider the data set on green buildings in greenbuildings.csv. This contains data on 7,894 commercial rental properties from across the United States. Of these, 685 properties have been awarded either LEED or EnergyStar certification as a green building. Here is a list of the variables:

- CS.PropertyID: the building's unique identifier in the database.
- cluster: an identifier for the building cluster, with each cluster containing one green-certified building and at least one other non-green-certified building within a quarter-mile radius of the cluster center.
- size: the total square footage of available rental space in the building.
- empl.gr: the year-on-year growth rate in employment in the building's geographic region.
- Rent: the rent charged to tenants in the building, in dollars per square foot per calendar year.
- leasing.rate: a measure of occupancy; the fraction of the building's available space currently under lease.
- stories: the height of the building in stories.
- age: the age of the building in years.
- renovated: whether the building has undergone substantial renovations during its lifetime.
- class.a, class.b: indicators for two classes of building quality (the third is Class C). These are relative classifications within a specific market. Class A buildings are generally the highest-quality properties in a given market. Class B buildings are a notch down, but still of reasonable quality. Class C buildings are the least desirable properties in a given market.
- green.rating: an indicator for whether the building is either LEED- or EnergyStar-certified.
- LEED, Energystar: indicators for the two specific kinds of green certifications.
- net: an indicator as to whether the rent is quoted on a "net contract" basis. Tenants with net-rental contracts pay their own utility costs, which are otherwise included in the quoted rental price.
- amenities: an indicator of whether at least one of the following amenities is available on-site: bank, convenience store, dry cleaner, restaurant, retail shops, fitness center.
- cd.total.07: number of cooling degree days in the building's region in 2007. A degree day is a measure of demand for energy; higher values mean greater demand. Cooling degree days are measured relative to a baseline outdoor temperature, below which a building needs no cooling.
- hd.total07: number of heating degree days in the building's region in 2007. Heating degree days are also measured relative to a baseline outdoor temperature, above which a building needs no heating.
- total.dd.07: the total number of degree days (either heating or cooling) in the building's region in 2007.

- Precipitation: annual precipitation in inches in the building's geographic region.
- Gas.Costs: a measure of how much natural gas costs in the building's geographic region.
- Electricity.Costs: a measure of how much electricity costs in the building's geographic region.
- City_Market_Rent: a measure of average rent per square-foot per calendar year in the building's local market.

Your goal is to build the best predictive model possible for *revenue per square foot per calendar year*, and to use this model to quantify the average change in rental income per square foot (whether in absolute or percentage terms) associated with green certification, holding other features of the building constant. Note that revenue per square foot per year is the product of two terms: rent and leasing_rate! This reflects the fact that, for example, high-rent buildings with low occupancy may not actually bring in as much revenue as lower-rent buildings with higher occupancy.

You can choose whether to consider LEED and EnergyStar separately or to collapse them into a single "green certified" category. You can use any modeling approaches in your toolkit (regression, variable selection, trees, etc), and you should also feel free to do any feature engineering you think helps improve the model. Just make sure to explain what you've done.

Write a short report detailing your methods, modeling choice, and conclusions, following the report-writing guidelines posted on the website.

Report

#Overview: We are give the task of determining how revenue per square foot per calendar year changes for apartments buildings that are green certified, holding other features constant. The idea is that green certified buildings should be more desirable as they save money in terms of utilities for the tenants and save the landlords money in basic electricity costs as well.

#Data and the model: We are looking at data that contains data on 7,894 commercial rental properties from across the United States. Of these, 685 properties, or about 8.6%, have been awarded either LEED or EnergyStar certification as a green building.

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.0     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

## naivebayes 0.9.7 loaded

##
## Attaching package: 'foreach'

## The following objects are masked from 'package:purrr':
##
##     accumulate, when
```

```
## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift

## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2

##
## The 'mosaic' package masks several functions from core packages in order to add
## additional features.  The original behavior of these functions should not be affected by this.

##
## Attaching package: 'mosaic'

## The following object is masked from 'package:Matrix':
##
##     mean

## The following object is masked from 'package:caret':
##
##     dotPlot

## The following object is masked from 'package:modelr':
##
##     resample

## The following objects are masked from 'package:dplyr':
##
##     count, do, tally

## The following object is masked from 'package:purrr':
##
##     cross

## The following object is masked from 'package:ggplot2':
##
##     stat

## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var

## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double()
## )
## i Use `spec()` for the full column specifications.
```

First, I will create a variable 'revenue.' This variable will represent the revenue per square foot per year that a building could obtain. I will generate it by multiplying the rent per square foot per year by the building leasing rate.

Now, I will split the data into a training and testing set. I will also fit a linear regression model that includes all the possible variables as predictors for revenue to ensure all variable are controlled for.

```
##       (Intercept)     CS_PropertyID           cluster              size
##             -2261                 0                 0                 0
##           empl_gr              Rent      leasing_rate           stories
##                 0                93                25                -1
##               age         renovated           class_a           class_b
##                 1               -18               -22               -14
##      green_rating              LEED        Energystar               net
##               -86                48                83               -17
##         amenities       cd_total_07       hd_total07      total_dd_07
##               -10                 0                 0                NA
##     Precipitation         Gas_Costs Electricity_Costs      cluster_rent
##                 2             -9610              2914                -5
```

```
## Warning in predict.lm(model, data): prediction from a rank-deficient fit may be
## misleading
```

```
## [1] 232.702
```

As we can see above there are some variables that have zero use reported.. We will remove these variables and re-run the linear regression from above.

```
##       (Intercept)              Rent      leasing_rate               age
##             -2172                93                25                 0
##         renovated           class_a           class_b      green_rating
##               -20               -29               -16               -83
##              LEED        Energystar               net         amenities
##                41                83                -2                -2
##     Precipitation         Gas_Costs Electricity_Costs      cluster_rent
##                 2             -1453              -155                -6
```

```
## [1] 234.2517
```

The RMSE decreased only slightly, likely due to variance in the data set. Moving forward, we will utilize this model and perform various feature engineering techniques in order to enhance the model.

First, I will begin by interacting the age and renovated variables.

```
##       (Intercept)              Rent      leasing_rate               age
##             -2202                93                25                 1
##         renovated           class_a           class_b      green_rating
##                56               -24               -12               -79
##              LEED        Energystar               net         amenities
##                54                81                -3                -3
##     Precipitation         Gas_Costs Electricity_Costs      cluster_rent
##                 2             -1228              -170                -6
##     age:renovated
##                -1
```

```
## [1] 234.2152
```

As we can see, this interaction has made the model worse off. We will remove this interaction and now focus on interacting the amenities and electricity costs variables.

```
##                  (Intercept)                         Rent
##                        -2106                           93
##                leasing_rate                          age
##                          25                            0
##                    renovated                      class_a
##                         -19                          -36
##                     class_b                 green_rating
##                         -19                          -94
##                        LEED                   Energystar
##                          52                           92
##                         net                    amenities
##                           1                         -190
##               Precipitation                    Gas_Costs
##                           2                         1349
##           Electricity_Costs                 cluster_rent
##                       -2990                           -6
## amenities:Electricity_Costs
##                        6173
```

```
## [1] 233.3581
```

This interaction improved the model by about 2 points. We will now look into a few more interactions to see if the model can be fit any better.

```
##                  (Intercept)                         Rent
##                        -2060                           90
##                leasing_rate                          age
##                          25                            1
##                    renovated                      class_a
##                         -19                          -33
##                     class_b                 green_rating
##                         -14                          -49
##                        LEED                   Energystar
##                          21                           49
##                         net                    amenities
##                          10                         -207
```

```
##               Precipitation                     Gas_Costs
##                           2                         -2945
##           Electricity_Costs                  cluster_rent
##                        -760                            -6
##              Rent:amenities amenities:Electricity_Costs
##                           6                          1139
```

```
## [1] 231.0283
```

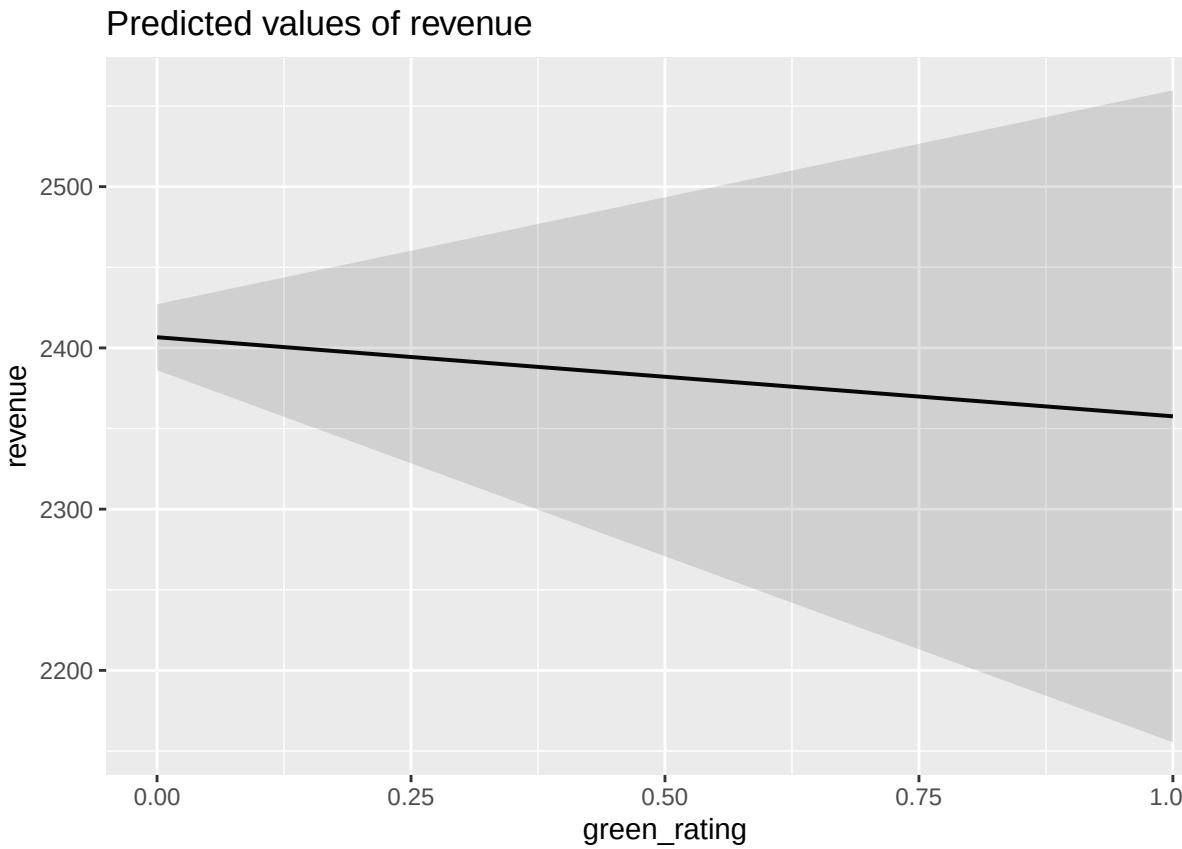
The interaction between amenities and electricity costs decreases the RMSE by about 6 more points. This is a fairly large change at this level.

#Results As we can see, the RMSE decreased each model until it resulted with an RMSE of 309.1046. This means that this linear model is the best prediction of revenue per year for the different apartment buildings.

If we calculate the marginal effects of our model, one can see that the change in the rent and leasing rate is relatively the same across the model.

```
##   dydx_Rent dydx_leasing_rate  dydx_age dydx_renovated dydx_class_a
## 1  95.75576           25.0582 0.5179797       -19.2418    -32.60334
## 2  95.75576           25.0582 0.5179797       -19.2418    -32.60334
## 3  95.75576           25.0582 0.5179797       -19.2418    -32.60334
## 4  89.71504           25.0582 0.5179797       -19.2418    -32.60334
## 5  95.75576           25.0582 0.5179797       -19.2418    -32.60334
## 6  95.75576           25.0582 0.5179797       -19.2418    -32.60334
##   dydx_class_b dydx_green_rating dydx_LEED dydx_Energystar dydx_net
## 1    -14.48771         -49.05623  21.08349        49.31526 10.46473
## 2    -14.48771         -49.05623  21.08349        49.31526 10.46473
## 3    -14.48771         -49.05623  21.08349        49.31526 10.46473
## 4    -14.48771         -49.05623  21.08349        49.31526 10.46473
## 5    -14.48771         -49.05623  21.08349        49.31526 10.46473
## 6    -14.48771         -49.05623  21.08349        49.31526 10.46473
##   dydx_amenities dydx_Precipitation dydx_Gas_Costs dydx_Electricity_Costs
## 1      58.606045           1.900076       -2944.68               378.4931
## 2      -1.690073           1.900076       -2944.68               378.4931
## 3      26.942963           1.900076       -2944.68               378.4931
## 4      37.151788           1.900076       -2944.68              -760.1004
## 5      71.523512           1.900076       -2944.68               378.4931
## 6      86.444103           1.900076       -2944.68               378.4931
##   dydx_cluster_rent
## 1         -5.654988
## 2         -5.654988
## 3         -5.654988
## 4         -5.654988
## 5         -5.654988
## 6         -5.654988
```

Now I will plot the predicted values of revenue based on whether or not a building is either LEED certified or



EnergyStar certified.
From the plot above, one can see that as the green rating goes from 0 to 1, the predicted revenue value decreases from slightly above 2400 to slightly below 2350. This would indicate that a building that is either LEED certified or EnergyStar certified does not create a higher renal income per square foot.

#Conclusion In conclusion, this model displayed that a building that is either LEED certified or EnergyStar certified does not create a higher renal income per square foot. One explanation could be due to the initial investment, landlords had to increase the rental rates in the green buildings. Due to the increase in these rental rates, tenants were more likely to choose a building that was not green certified. This means the tenants value the rental rate over being green so to speak.

Predictive model building: California housing

The data in `CAhousing.csv` containts data at the census-tract level on residential housing in the state of California. Each row is a census tract, and the columns are as follows:

- longitude, latitude: coordinates of the geographic centroid of the census tract
- housingMedianAge: median age in years of all residential households in the census tract
- population: total population of the tract
- households: total number of households in the tract.
- totalRooms, totalBedrooms: total number of rooms and bedrooms for households in the tract. NOTE: these are *totals*, not averages. Consider standardizing by households.

- medianIncome: median household income in USD for all households in the tract.
- medianHouseValue: median market value of all households in the tract.

Your task is to build the best predictive model you can for `medianHouseValue`, using the other available features. Write a short report detailing your methods. Make sure your report includes an estimate for the overall out-of-sample accuracy of your proposed model. Also include three figures:

- a plot of the original data, using a color scale to show medianHouseValue (or log medianHouseValue) versus longitude (x) and latitude (y).
- a plot of your model's predictions of medianHouseValue (or log medianHouseValue) versus longitude (x) and latitude (y).
- a plot of your model's errors/residuals (or log residuals) versus longitude (x) and latitude (y).

You can get nearly full credit (but not 100%) without a mapping package, i.e. just treating longitude and latitude as generic x/y coordinates. But a modest number of points will be reserved for those who can successfully show these plots in a visually pleasing fashion on an *actual map of California*. This will entail learning how to use an R package capable of making maps. (We haven't covered this in class, but a major part of being a data scientist is learning how to use new software tools and libraries "on the fly" like this.) I recommend `ggmap` as a good starting point, but you can use whatever R tools you want here.

#Overview I will be attempting to predict median house value in various areas of the state of California. The data used has several factor variables that can contribute to a houses value.

#Data and Model I used data from a census-track level for housing in the state of California.

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   longitude = col_double(),
##   latitude = col_double(),
##   housingMedianAge = col_double(),
##   totalRooms = col_double(),
##   totalBedrooms = col_double(),
##   population = col_double(),
##   households = col_double(),
##   medianIncome = col_double(),
##   medianHouseValue = col_double()
## )
```

Below is a map of the state of California with the median house value mapped. We can see that the median housing price increases the further toward the coast a house is.

```
## Google's Terms of Service: https://cloud.google.com/maps-platform/terms/.
```

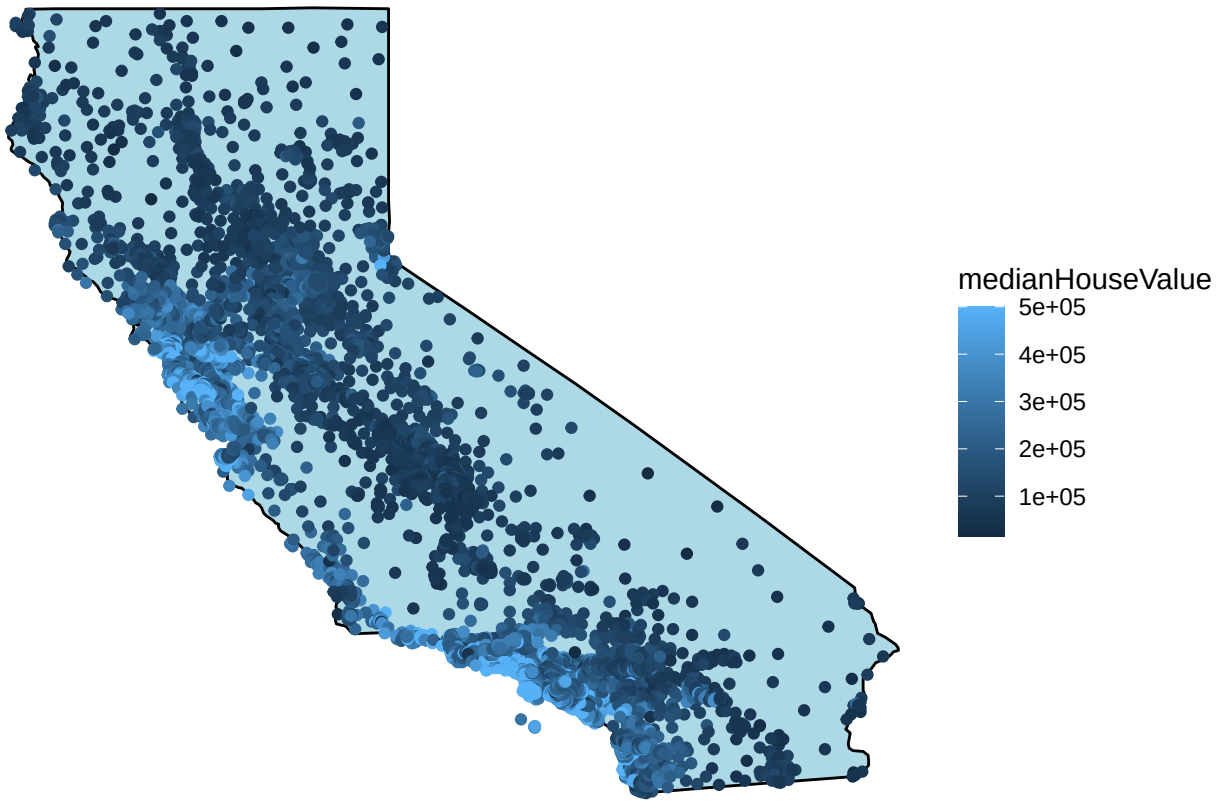
```
## Please cite ggmap if you use it! See citation("ggmap") for details.
```

```
##
## Attaching package: 'maps'
```

```
## The following object is masked from 'package:purrr':
##
##     map
```

```
## Loading required package: viridisLite
```

```
##          long      lat group order     region subregion
## 667 -120.0060 42.00927     4   667 california      <NA>
## 668 -120.0060 41.20139     4   668 california      <NA>
## 669 -120.0060 39.70024     4   669 california      <NA>
## 670 -119.9946 39.44241     4   670 california      <NA>
## 671 -120.0060 39.31636     4   671 california      <NA>
## 672 -120.0060 39.16166     4   672 california      <NA>
```



Now I will begin to model the home prices to attempt to develop the best model for predictive behavior.

First, the total bedrooms and total rooms variables are for an entire region, not per household. We need to divide each by households to standardize the variables

```
housing$rooms <- housing$totalRooms / housing$households
housing$bedrooms <- housing$totalBedrooms / housing$households
```

Next, we can begin modeling. I will first split the data into a training set and a testing set of data. Then I will fit a linear regression model to the training data. I will start with an extremely basic model that utilizes the variables longitude, latitude, and households to predict the median house value.

```
## (Intercept)        long         lat  households
##    -5898822      -71891      -70071          14
```

```
## [1] 99708.7
```

One can see above the basic model reports an extremely high RMSE. I will now work on improving the model through feature engineering. First to improve the model, I will add the variables for housing median age, median income, and rooms.

```
##      (Intercept)             long              lat       households
##         -3921198           -46327           -45351               27
##            rooms     medianIncome housingMedianAge
##             2558            36564             1243
```

```
## [1] 71577.33
```

We can see already the RMSE has been greatly reduced just by the introduction of a few extra variables. I will add bedrooms to see if the RMSE can be reduced further.

```
##      (Intercept)             long              lat       households
##         -3754672           -43668           -42049               25
##            rooms     medianIncome housingMedianAge         bedrooms
##           -10567            43558             1207            72981
```

```
## [1] 72091.14
```

One can see adding bedrooms reduces the RMSE, but by only about 350. While this is still an improvement, it is not significant. There is one last variable, population, we can add to the basic model. I originally theorized that this variable would cause colinearity in the model so I intentionally omitted it. I will include this variable now to test if my theory is correct.

```
##      (Intercept)             long              lat       households
##         -3649170           -43189           -42951              148
##            rooms     medianIncome housingMedianAge         bedrooms
##            -7257            41604             1129            54875
##       population
##              -46
```

```
## [1] 70466.67
```

My theory was proven incorrect, and by including population the RMSE was reduced by about 3000 points. Moving forward, I will include it in the model.

Now I will square the previous linear model that includes all variables to see if it will improve the RMSE.

```
##                   (Intercept)                          long
##                        -75419                          6452
##                           lat                    households
##                         88255                         -1994
##                         rooms                  medianIncome
##                        306770                       -816760
```

```
##              housingMedianAge                      bedrooms
##                       -100901                     -1736253
##                    population                     long:lat
##                           495                           523
##               long:households                   long:rooms
##                           -27                          4411
##             long:medianIncome        long:housingMedianAge
##                        -10640                         -1209
##                long:bedrooms              long:population
##                        -22862                             8
##                lat:households                   lat:rooms
##                           -35                          6266
##             lat:medianIncome         lat:housingMedianAge
##                        -11943                         -1244
##                lat:bedrooms               lat:population
##                        -28622                            11
##             households:rooms      households:medianIncome
##                             3                            17
##   households:housingMedianAge          households:bedrooms
##                             7                          -105
##         households:population           rooms:medianIncome
##                             0                          -591
##        rooms:housingMedianAge               rooms:bedrooms
##                          -153                          -114
##             rooms:population medianIncome:housingMedianAge
##                            -5                            68
##         medianIncome:bedrooms      medianIncome:population
##                          5814                            -1
##     housingMedianAge:bedrooms   housingMedianAge:population
##                          1384                            -2
##           bedrooms:population
##                            83
```

```
## [1] 65617.42
```

One can see that the RMSE was improved by about 300 points.

Now we can try several different feature engineering methods to potentially improve the model further. We will begin by exploring whether interacting two variables can make a difference.

```
##                            (Intercept)
##                                 165366
##                                   long
##                                   8906
##                                    lat
##                                  88191
##                             households
##                                  -3035
##                                  rooms
##                                 312663
##                           medianIncome
##                                -908972
##                        housingMedianAge
##                                 -99024
```

```
##                                bedrooms
##                                -1677134
##                              population
##                                     619
##                                long:lat
##                                     515
##                           long:households
##                                     -39
##                               long:rooms
##                                    4459
##                          long:medianIncome
##                                  -11648
##                     long:housingMedianAge
##                                   -1190
##                             long:bedrooms
##                                  -22288
##                           long:population
##                                       9
##                            lat:households
##                                     -43
##                                lat:rooms
##                                    6398
##                           lat:medianIncome
##                                  -12590
##                        lat:housingMedianAge
##                                   -1228
##                               lat:bedrooms
##                                  -28562
##                             lat:population
##                                      12
##                           households:rooms
##                                      -9
##                      households:medianIncome
##                                     206
##                 households:housingMedianAge
##                                       6
##                        households:bedrooms
##                                    -134
##                      households:population
##                                       0
##                         rooms:medianIncome
##                                   -1763
##                      rooms:housingMedianAge
##                                    -169
##                            rooms:bedrooms
##                                    -195
##                         rooms:population
##                                      -5
##               medianIncome:housingMedianAge
##                                      19
##                      medianIncome:bedrooms
##                                    7749
##                   medianIncome:population
##                                       0
```

```
##                housingMedianAge:bedrooms
##                                     1466
##             housingMedianAge:population
##                                       -2
##                      bedrooms:population
##                                       83
##            long:households:medianIncome
##                                        2
##             lat:households:medianIncome
##                                        1
##           households:rooms:medianIncome
##                                        3
## households:medianIncome:housingMedianAge
##                                        0
##         households:medianIncome:bedrooms
##                                        7
##       households:medianIncome:population
##                                        0
```

```
## [1] 65471.31
```

We can see an improvement of about 38 points. Let's try to interact a few other variables.

```
##                               (Intercept)
##                                    701916
##                                      long
##                                     12172
##                                       lat
##                                     84481
##                                households
##                                     -3511
##                                     rooms
##                                    154418
##                              medianIncome
##                                 -1150457
##                           housingMedianAge
##                                    -82180
##                                  bedrooms
##                                 -1546169
##                                population
##                                       702
##                                  long:lat
##                                       507
##                            long:households
##                                       -40
##                                 long:rooms
##                                      2995
##                           long:medianIncome
##                                    -13562
##                         long:housingMedianAge
##                                      -948
##                                long:bedrooms
##                                    -21459
```

```
##                          long:population
##                                        9
##                           lat:households
##                                      -42
##                                lat:rooms
##                                     5417
##                           lat:medianIncome
##                                   -12810
##                       lat:housingMedianAge
##                                     -921
##                                lat:bedrooms
##                                   -27613
##                           lat:population
##                                       12
##                          households:rooms
##                                       31
##                     households:medianIncome
##                                      386
##                households:housingMedianAge
##                                        8
##                         households:bedrooms
##                                      -70
##                        households:population
##                                        0
##                         rooms:medianIncome
##                                    44890
##                     rooms:housingMedianAge
##                                      349
##                            rooms:bedrooms
##                                      353
##                           rooms:population
##                                      -18
##               medianIncome:housingMedianAge
##                                    -3343
##                        medianIncome:bedrooms
##                                    -8610
##                      medianIncome:population
##                                      -26
##                   housingMedianAge:bedrooms
##                                     -909
##                 housingMedianAge:population
##                                       -3
##                         bedrooms:population
##                                       55
##               long:households:medianIncome
##                                        3
##          long:medianIncome:housingMedianAge
##                                      -54
##                    long:rooms:medianIncome
##                                      431
##                 lat:households:medianIncome
##                                        2
##           lat:medianIncome:housingMedianAge
##                                      -74
```

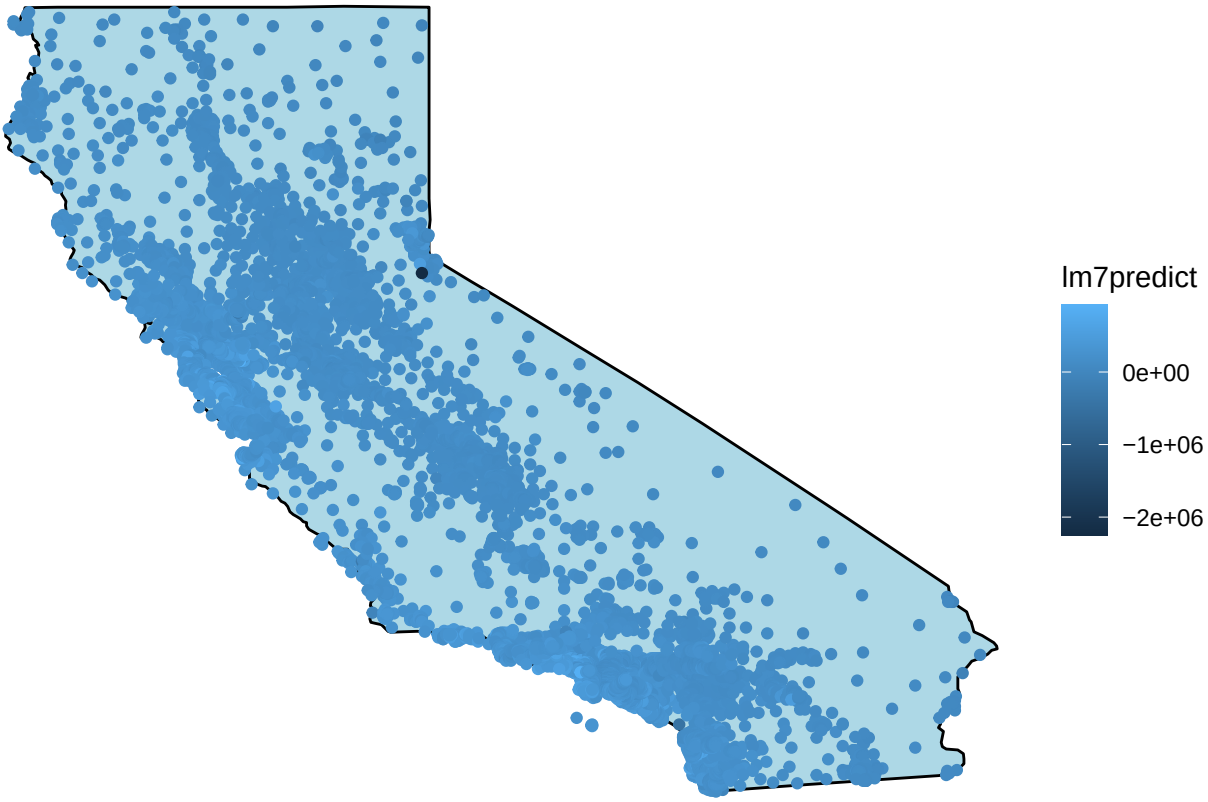
```
##                   lat:rooms:medianIncome
##                                      294
## households:medianIncome:housingMedianAge
##                                       -1
##            households:rooms:medianIncome
##                                       -8
##      rooms:medianIncome:housingMedianAge
##                                     -166
##         households:medianIncome:bedrooms
##                                        3
##   medianIncome:housingMedianAge:bedrooms
##                                      731
##              rooms:medianIncome:bedrooms
##                                     -262
##       households:medianIncome:population
##                                        0
## medianIncome:housingMedianAge:population
##                                        0
##            rooms:medianIncome:population
##                                        3
```

```
## [1] 75295.81
```

One can see that by adding a few other interaction variables, we were able to achieve the lowest RMSE yet.

#Results Let's go ahead and re-plot the original map but now with the new predicted values from the best predicting linear model.

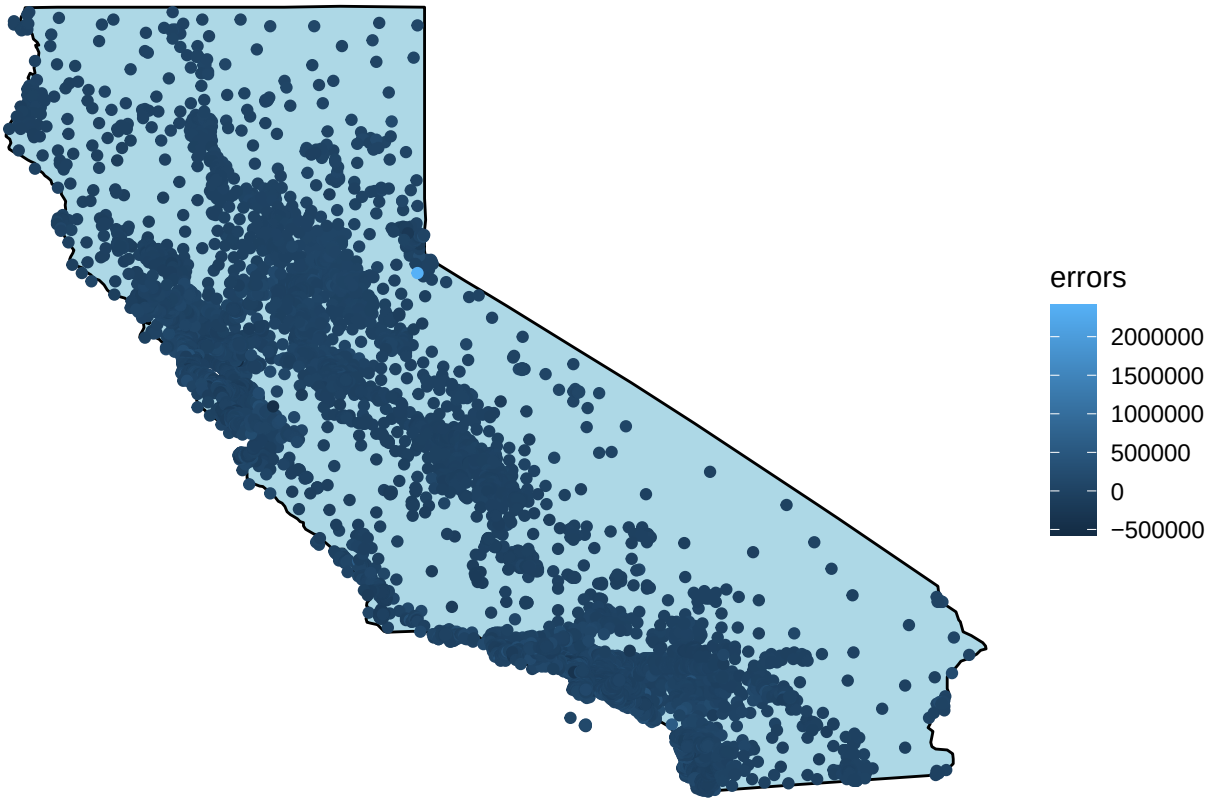
```
##          long      lat group order     region subregion
## 667 -120.0060 42.00927     4   667 california      <NA>
## 668 -120.0060 41.20139     4   668 california      <NA>
## 669 -120.0060 39.70024     4   669 california      <NA>
## 670 -119.9946 39.44241     4   670 california      <NA>
## 671 -120.0060 39.31636     4   671 california      <NA>
## 672 -120.0060 39.16166     4   672 california      <NA>
```



It appears our model does a good job of predicting a similar result to those of the original plot. One can see that the houses as before are more expensive near the coast than they ar inland. there is an area at the elbow of the east side of the state that is showing similar values to that of the original plot.

I will now plot the errors over the state of California. Here I defined errors as the difference between the actual values for median house value and the predicted values for the seventh linear model.

```
##          long      lat group order     region subregion
## 667 -120.0060 42.00927     4   667 california      <NA>
## 668 -120.0060 41.20139     4   668 california      <NA>
## 669 -120.0060 39.70024     4   669 california      <NA>
## 670 -119.9946 39.44241     4   670 california      <NA>
## 671 -120.0060 39.31636     4   671 california      <NA>
## 672 -120.0060 39.16166     4   672 california      <NA>
```



The errors are centered around zero, indicating our model does a good job of predicting the median house value.

#Conclusion In conclusion the linear model I built did a good job of predicting the median house value for a home in the state of California. Based on the lowest RMSE possible and a map of the errors that would indicate most are around zero, the model predicted similar values for median housing prices to that of the original data.