Data Mining: Exercises #3

Gaetano Dona-Jehan and Jordan Fox

Question #1: Policing and Crime

1) Why can't we regress crime onto police?

The direct effect of policing on crime is difficult to estimate without an instrument due to the fact that the amount of police on the streets is endogenous to the crime level. Essentially, the researchers are facing a problem of reverse causality: high crime rates lead to more police on the beat. This makes naive regressions of crime onto police levels inherently biased, as the level of police is correlated with the unobservables, a violation of one of the key assumptions for ordinary least-squares.

2) How do the researchers isolate this effect?

The researchers are able to isolate the effect of policing on crime by using the terror threat level in DC as an instrument for the amount of police on patrol. This instrument satisfies the exclusion restriction, in that terror alerts don't impact the crime rate *except* through an increased number of police on the ground. They use an indicator variable, equal to 1 on days where the terror threat level was on 'high'.

The first column shows the effect of the high terror alert on crime, without controlling for the ridership on major public transit in DC. The impact of a terror alert was a 7.316 unit decrease in crime; this effect was significant at the 5% level. The second column controls for ridership, as it's possible that the high terror alerts decreases the number of potential victims for criminals to target and that this is the channel that high terror alerts decrease the crime rate. However, we can see that the impact of the alerts, when controlling for ridership, is still a decrease in crime by just over 6 units, significant at the 5% level. The R-square for column 1 is .15, whereas for column 2 this is .17, indicating that these models describe about 15% and 17% of the variation in the data, respectively.

3) Why control for metro ridership?

The researchers control for ridership because increased ridership may be a significant determinant of crime levels; more riders on public transit, particularly in areas around the capitol, may correlate with higher crime rates because an increase in the number of tourists might embolden criminals to target these areas. Controlling for ridership in this case prevents any variation in crime levels caused by the change in ridership from being attributed to the high terror alert, giving us a less-biased coefficient on our crime variable.

4) Describing the model in Table 4

The model in table 4 allows for the effect of the high terror alert on crime to vary by district. It does so by interacting "high alert" with two binary variables: one for District 1, and a second for districts outside of District 1. It's reasonable to expect the effect of the alert to have a different effect on districts outside of District 1. We might expect more police to be on patrol in District 1 because it's the site of US Capitol, and thus we can expect most of the police to be deployed here in the event of an alert.

Based on the output, a high alert is associated with a decrease in crime by 2.62 units within District 1. For other districts, this decrease is only by .5 units. We don't see an R-squared reported, so we can't tell if this model performs better than column two in Table 2.

Question #2: Green Buildings Certification Prediction

For this question, we are considering the data called green buildings, which contains data on 7,894 commercial rental properties from across the United States. Of these, only 685 properties are considered as "green" properties, as they were awarded either LEED or EnergyStar certification. The goal here is to build the best predictive model possible for revenue per square foot per calendar year. Once we have our best predictive model possible for revenue per square foot per calendar year. Once we have our best predictive models, we will use them to quantify the average change in rental income per square foot associated with green certification, assuming that there is a change, holding all other features of the building constant.

The first step is to generate the variable revenue per square foot per year, which we know is the produce of the two terms: rent and leasing_rate. We can confirm that these variables exist by using the head() function:

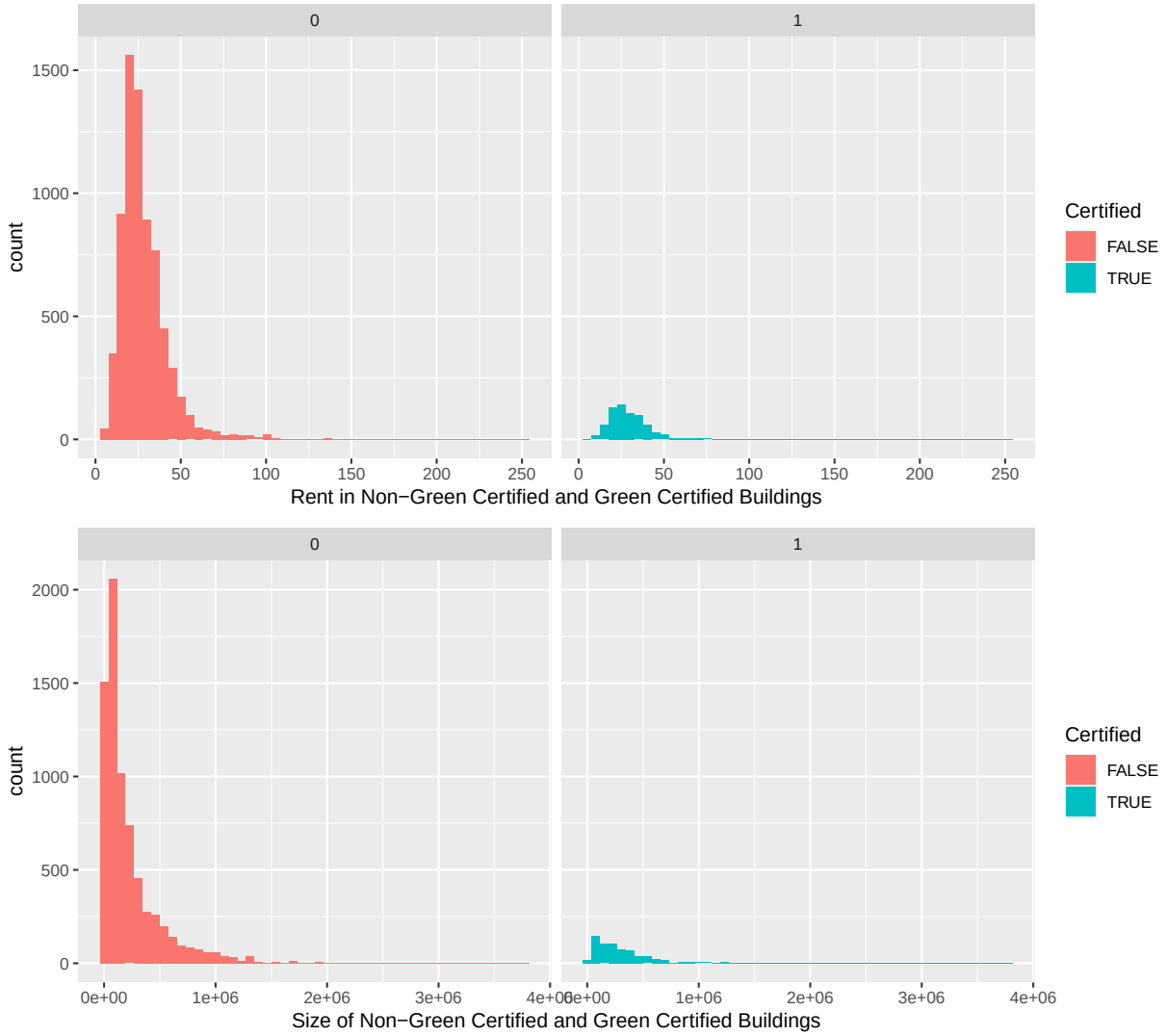
```
##   CS_PropertyID cluster   size empl_gr  Rent leasing_rate stories age renovated
## 1        379105       1 260300    2.22 38.56        91.39      14  16         0
##   class_a class_b LEED Energystar green_rating net amenities cd_total_07
## 1       1       0    0          1            1   0         1        4988
##   hd_total07 total_dd_07 Precipitation Gas_Costs Electricity_Costs
## 1         58        5046         42.57    0.0137             0.029
##   City_Market_Rent  Revenue
## 1            36.78 3523.998
```

As we can see from the head command, we have successfully created the necessary revenue variable.

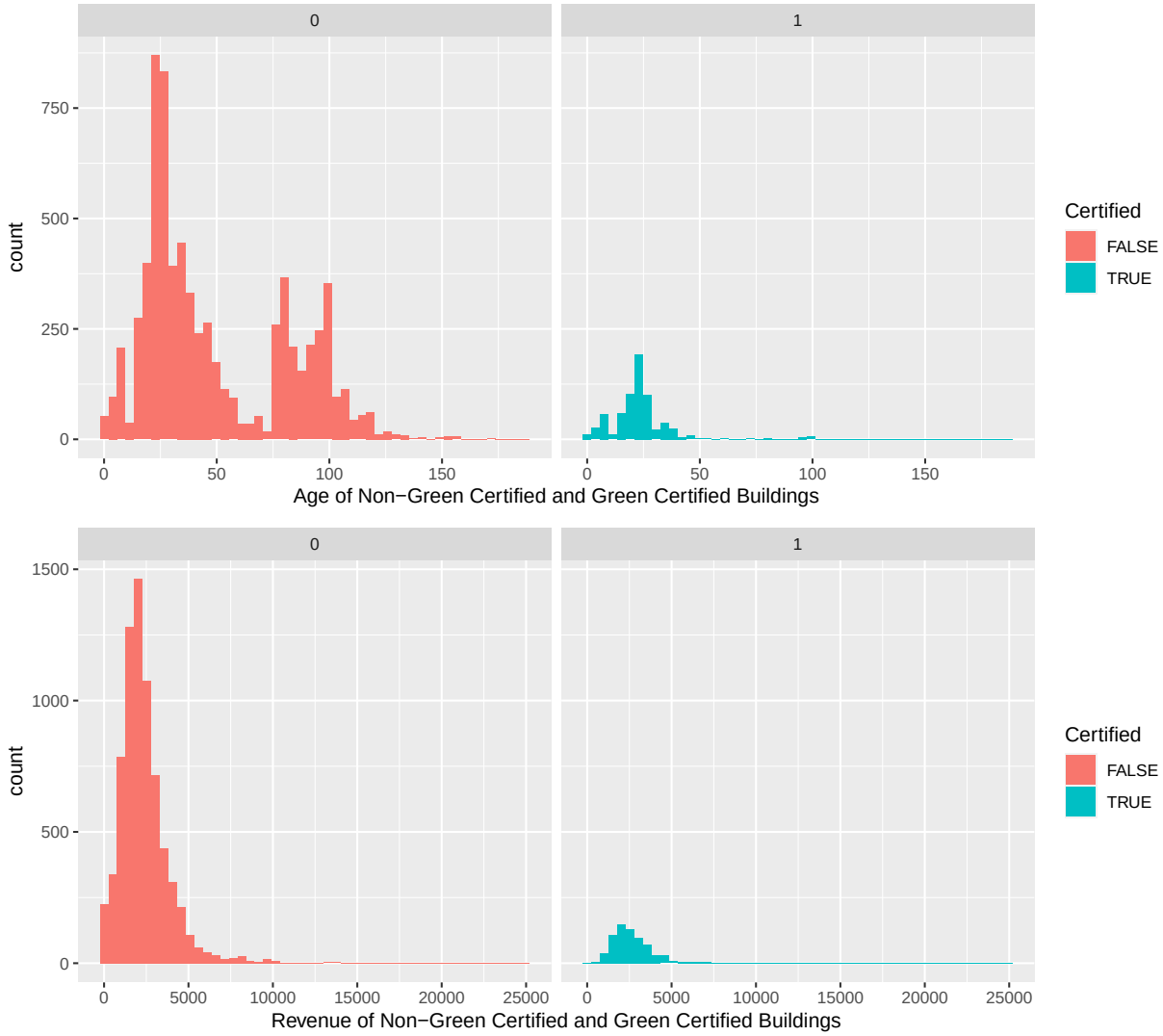
However, there are now a total of 24 variables in our Green dataset, and we would like to make things a bit more manageable. Consequently, we decided to get rid of the CS.PropertyID, cluster, Rent, leasing_rate, LEED and EnergyStar variables. The reason why we decided to remove rent and leasing_rate is due to the fact that they would be highly correlated with our revenue variable, as the revenue variable is a product of these two variables, so it would throw off our analysis. Since we are only looking at the impact of green certification, we decided to get rid of the LEED and Energystar variable because they seemed redundant in the precense of our green certified variable. Finally, we got rid of the PropertyID and the cluster simply because we do not believe that they are necessary for our model.

Green Buildings: Preliminary Analysis

Here, we plot some of the distributions of some of the variables in the data. While we don't use all of these in our analysis, they are nevertheless useful for trying to get an idea about the differences between certified and non-certified buildings.



Above, it's immediately obvious that non-certified buildings greatly outnumber the certified ones. Certified buildings appear to have a slightly higher average rent, but it's not clear whether certified or non-certified apartments have a greater average size. All of the distributions appear to have a right skew.



Next, we see that there appear to be two different groups of apartments in the non-certified group; those that were built around 100 years ago, and those build within the last 50 years. This leads to a near-multimodal appearance for the distribution. Meanwhile, the mean age of certified buildings appears to be about 25 years old. However, what is not apparent is whether or not green-certified buildings bring in greater revenues than non-green buildings. We'll proceed by building a model that helps us tease out the effect of being green-certified.

To create the model, we used two methods. The first method is forward stepwise selection. It is important to note that this is a computational heavy process, and depending on the machine running the code, it could be very time consuming. As a result, we do not include the output from the stepwise process.

Stepwise Selection Results

In-Sample RMSE:

```
## [1] 993.501
```

Out-of-Sample RMSE:

```
## [1] 975.9966
```

Although not shown for aesthetic reasons, our in-sample model predicts that, on average, a green-certified building will generate $411.25 per square foot. Out out-of-sample model predicts an effect of $342 per square foot.

Our stepwise model predicts that, on average, a green-certified building will generate $411.25 more than a non-certified building.

Now we will see if we get similar results using the cross validation for gamma lasso penalty selection.

Coefficients from Lasso Regression:

```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                            seg100
## intercept         -5.036618e+02
## size               5.752493e-05
## empl_gr            .
## Rent               9.334940e+01
## stories            4.034989e+00
## age                .
## renovated          .
## class_a            8.575012e+01
## class_b            5.190929e+01
## green_rating       5.589713e+01
## net                .
## amenities          9.086914e+01
## cd_total_07        .
## hd_total07         8.255804e-03
## total_dd_07        .
## Precipitation      .
## Gas_Costs          9.713632e+02
## Electricity_Costs  .
## City_Market_Rent   1.401222e+00
```

Lasso Regression's RMSE (mean RMSE across 20 folds):

```
## [1] 713.0663
```

Although we get different coefficients from our approaches, due to the different selection and analysis methods, we can conclude that the effect of a green certification on revenue will be as low as $139 and as high as $411 per square foot. We now discuss our reasoning for selecting which model we believe is more accurate.

Model Selection: Step Function vs Lasso

While our step function method produced a lower RMSE, we decide to go with the results from our lasso regression. The reason is two-fold: first, there are many interactions in our step function model that do not make intuitive sense. For example, there isn't any reason to think that the interaction between precipitation and gas costs would help predict revenue per square foot. These superfluous interactions might be pushing our RMSE down, but at the cost of over-fitting.

Second, the gulf between the predicted impact of green certification between the two models raises some concerns for the estimate from our step function method. We know that the lasso penalizes large beta coefficients, which explains this discrepancy. Additionally, the RMSE between the step function and lasso specifications are not radically different, given the gulf between the coefficient on green certification between the two models. Because of this, we err on the side of caution and go with the estimate from our lasso regression, which predicts an additional $139.35 of revenue per square foot.

Question #3: California Housing Predictive Model

We're asked to build the best model for predicting the median house value of a census tract, and then to plot the median home value, predicted median home value, and residual onto a map of the state of California. We'll use step() from the library stats to build our model, and then we'll use map_data() from ggplot2 to plot these onto maps of California to give our analysis a spatial dimension.

Description of Data

The data we're given is a set of about 20,000 observations of census tracts from the state of California, with columns on latitude/longitude, median house age, population, demeaned number of households, number of rooms, median income, and median house value of the census tract. We standardize the households column by demeaning it, which gets things onto an appropriate scale.

Model Selection

We used forward selection to determine the best model. Because forward selection is computationally demanding, we also have the block of code which calls step() set to eval= FALSE.

The results from our forward step selection indicate that the lowest AIC is given by a model of medianHouseValue that considers medianIncome, housingMedianAge, totalBedrooms, population, totalRooms, latitude, longitude, and d_households, along with a number of interactions. Many of these make intuitive sense. For example, we might expect the medianIncome to vary with the longitude, as the cost of living is higher along the coastline; likewise, interacting population with longitude is reasonable because while the value of a home increases as we get closer to the coast, but also as the population of the area increases (ie, being located in a city).

Prediction Error for Forward-Selected Models

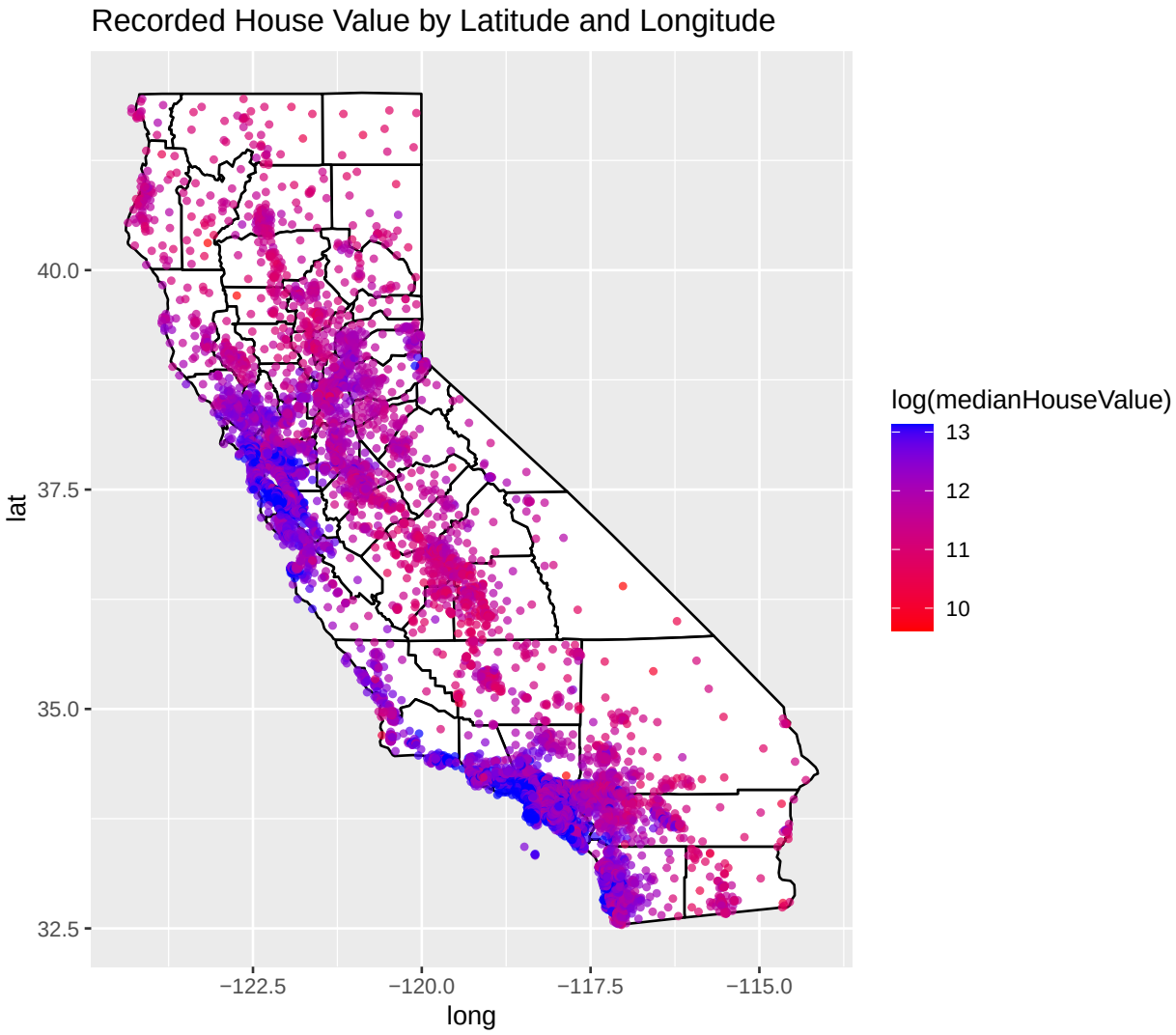
```
## [1] 0.3197307
```

```
## [1] 0.3161991
```

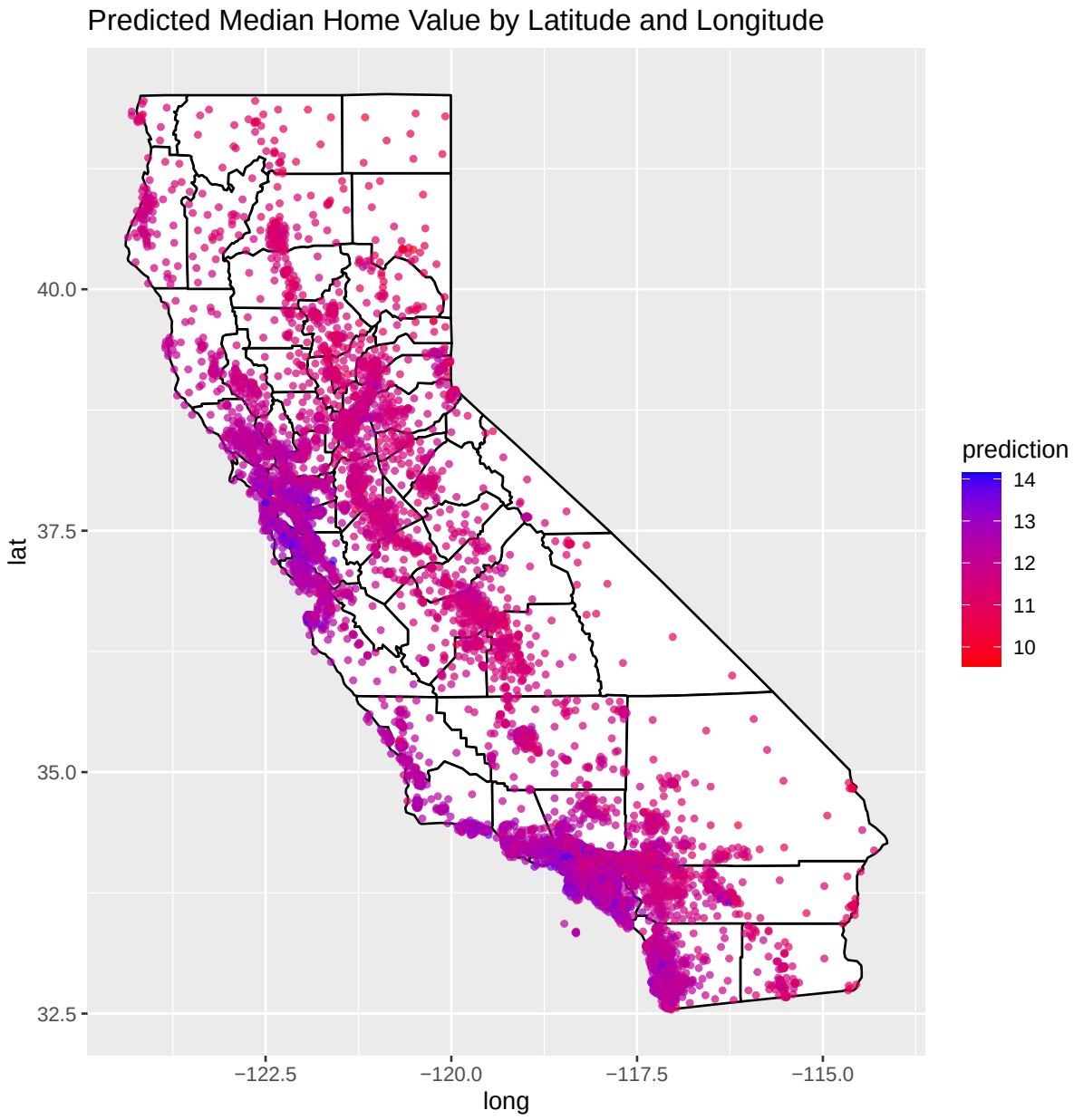
The RMSE is .3197307 and .3161991 for the in-sample and out-of-sample predictions, respectively. The transformation of the dependent variable into the log of median home value means that our predictions are off by about 31%, on average. For a home with a value of $100,000, we could thus reasonably expect a prediction to be somewhere in the range of $69,000 to $131,000. Below, we plot the median home value, the predicted median home value, and the prediction error by latitude and longitude.

Figures: Median Income, Predicted Income, and Prediction Error versus Census Tract in California

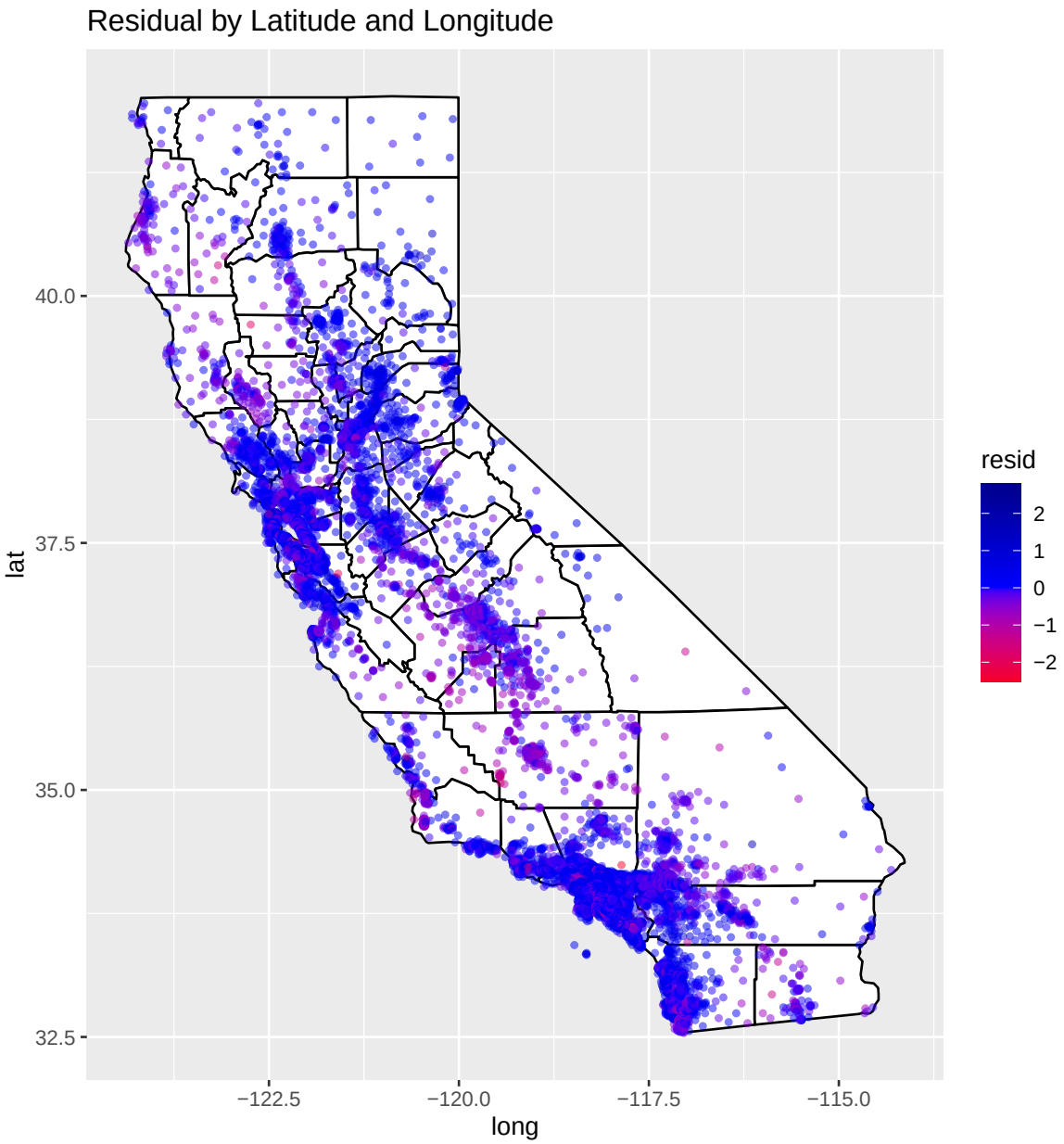
Note: We chose *red* = *low* and *blue* = *high* in our color gradient. Although this runs against the fact that people generally associate red with heat and blue with cold, ultimately it makes our figures easier on the eyes and makes the gradient along longitude in our first two figures easier to see.



The figure above shows us the medianHouseValue plotted against latitude and longitude. We can observe a gradient of sorts along longitude, indicating that home values are increasing as they get closer to the coast. Additionally, observations in and around cities have higher home values.



Here, we have the predicted median home value plotted against latitude and longitude. Our plot shows a similar gradient as the previous figure, with the home values along the coastline having a higher predicted value. The highest-predicted home values appear to be in Los Angeles and the Bay Area.



Above, we have the the residual for each predicted value plotted against latitude and longitude. Here, we don't see a gradient along longitude as we did in the previous figures; over-estimations and under-estimations appear to be evenly distributed throughout the plot.