Exercise 3

Question 1.1

Taking the data from a handful of cities would only establish correlation and not causation. To derive a casual relationship between number of police and the amount of crime, there would needed to be data taken of one variable that is unrelated to the other variable.

Question 1.2

The researchers are UPenn were able to isolate the effect by using the Terror Alert System and data from Washington, D.C. Since Washington, D.C. is considered a high target for terrorism, as the threat level increased to orange on the Terror Alert System, more police officers are added around the city, regardless of crime. Researchers were then able to observe a causal relationship between police and crime.

At the 5% level of significance, daily crime decreases by 7.316 occurrences when the Terror Alert System is high, all else equal. When controlling for midday ridership, daily crime decreases by 6.046 occurrences, all else equal. Additionally, at the 1% level of significance, midday ridership increases by 1734.1% when the Terror Alert System is high. In all cases, the R^2 is close to zero, which would be expected since police officers were added as a result of a high terrorist threat, and not in response to crime.

Question 1.3

When controlling for Metro ridership, the researchers were attempting to capture the effect on tourism in Washington, D.C. when the terror threat level was orange so as to understand if there would be fewer victims in a terrorist attack.

Question 1.4

The model being estimated here is the dependent variable of daily total number of crimes regressed on two interaction variables and a log taken for midday ridership. The two interactions are both interacting with the variable High Alert, with District 1 referring to a dummy variable.

All else equal, daily total number of crimes decreases by 2.621 occurrences in District 1 when the terror threat level is high. All else equal, daily total number of crime decreases by 0.571 occurrences in other districts when the terror threat level is high. All else equal, midday ridership increases 247.7% when the terror threat level is high. All else equal, daily total number of crimes decreases by 11.058 occurrences when the terror threat level is high.

The conclusion is that a high terror alert increases the number of police around the city, which subsequently also increases the number of Metro riders and decreases crime in all areas of Washington, D.C., most notably in District 1.

Question 2

I. Overview

The objective is to analyze a data set consisting of 7,894 commercial rental properties from across the United States and create a predictive model so as to determine the effect of green buildings on revenue per square foot per calendar year. Of the 7,894 properties, 685 properties are considered to be a green building. To best achieve this objective, variable selection was used to narrow the data to relevant buildings classified as green, while holding all over variables constant.

II. Data and Model

The creation of an interaction variable between rent and leasing_rate was used to predict the dependent variable, Annual_revenue_per_sqft. This varibale controls for any variance between high and low rent buildings and their appropriate occupancy rates throughout the year.

Buildings are classified as green through either the LEED or Energy Star certification. Within the data set, the dummy variable, green_rating, encompasses both certifications and was used to narrow the data to the relevant information. Of the resulting 685 properties considered to be green, 6 observations lacked necessary information, and were subsequently dropped from the data set, resulting in 679 total observations.

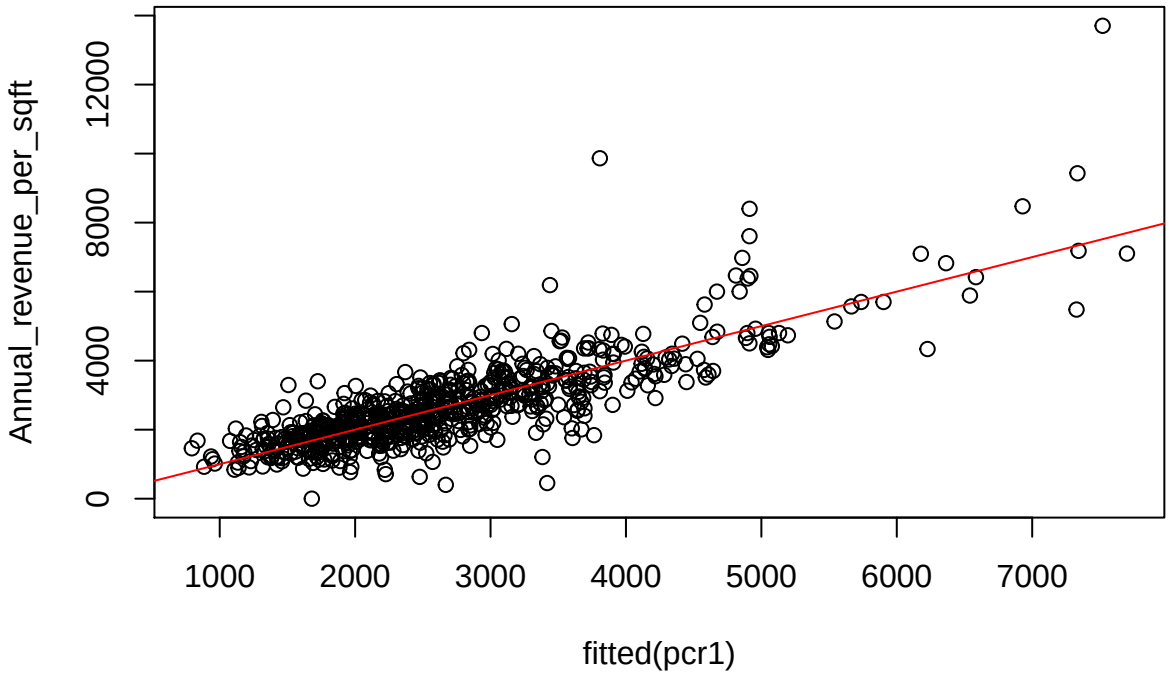
The next step is to use principal component analysis (PCA) as a means to fit a linear regression model to predict annual revenue per sqft. The goal of PCA is to use the remaining variables and create new "variables", called principal components (PCs). These PCs are scaled linear combinations of the original variables, but are also uncorrelated. There are 18 variables that were used to create the appropriate PCs, which include gas and electrical costs, and age of the building.

III. Results

```
##
## Call:
## lm(formula = y ~ greendata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2964.7  -379.6   -36.3   332.2  6183.2
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.701e+03  4.815e+01  56.093  < 2e-16 ***
## greendataPC1  -2.344e+02  7.721e+01  -3.036 0.002490 **
## greendataPC2   3.207e+02  3.887e+01   8.252 8.49e-16 ***
## greendataPC3  -8.965e+01  2.541e+01  -3.528 0.000447 ***
## greendataPC4  -1.752e+02  2.845e+01  -6.156 1.29e-09 ***
## greendataPC5   4.841e+01  4.946e+01   0.979 0.328077
## greendataPC6   2.411e+02  5.098e+01   4.729 2.76e-06 ***
## greendataPC7   3.176e+02  8.231e+01   3.859 0.000125 ***
## greendataPC8  -6.603e+01  8.819e+01  -0.749 0.454295
## greendataPC9  -2.535e+02  3.125e+01  -8.111 2.45e-15 ***
## greendataPC10 -5.839e+02  4.503e+01 -12.967  < 2e-16 ***
## greendataPC11  1.194e+02  4.753e+01   2.511 0.012263 *
## greendataPC12 -2.046e+02  1.312e+02  -1.560 0.119265
## greendataPC13  5.108e+02  4.899e+01  10.427  < 2e-16 ***
```

```
## greendataPC14  2.297e+02  6.023e+01   3.813 0.000150 ***
## greendataPC15 -4.199e+01  6.968e+01  -0.603 0.547024
## greendataPC16  1.394e+01  1.165e+02   0.120 0.904785
## greendataPC17  4.801e+01  1.732e+02   0.277 0.781755
## greendataPC18 -7.192e+16  3.939e+17  -0.183 0.855161
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 736.7 on 660 degrees of freedom
## Multiple R-squared:  0.6846, Adjusted R-squared:  0.676
## F-statistic: 79.59 on 18 and 660 DF,  p-value: < 2.2e-16
```

Figure 1: Green Building Revenue



Using PCA, I fit a linear regression model to predict annual revenue per sqft. Figure 1 displays a positive association between annual revenue per sqft and all other variables for green buildings. Most observations lie between 1000 and 4000 along the x-axis, with a predicted value of 1000-4000 in annual revenue per sqft. There also appears to be several outliers above the line of regression.

IV. Conclusion

Based on the results presented, there appears to be a positive associated between the amount of annual revenue per sqft that a green building can generate. Due to a buildings green certification, decreased gas and electrical costs may have an attractive sales appeal for perspective tenants, which may lead to higher prices that a building is able to charge. Moreover, clusters of certified green buildings would have a negative effect on prices, as it is direct competition in the area. To summarize, it would be a wise business objective for a building to invest and seek green certification.

Question 3

I. Overview

The objective is to analyze the data set collected by the census tract for residential housing in the state of California, and develop a predictive model that could accurately determine the median house value based on several features. This model could be helpful in determining a pattern of median house values in the state.

II. Data and Model

In this model, the variable medianHouseValue is the dependent variable and there are 8 independent variables, such as medianHouseIncome, housingMedianAge, and location, which is represented by longitude and latitude.

Two variables, totalRooms and totalBedrooms, are total numbers of rooms and bedrooms in a municipality. To make a better predictive model, both variables have been divided by the number of households of the municipality, so as to find the averages in each, and subsequently standardized.

The next step is to train, test, split the original data set, which then provides an appropriate testing data set for the predictive model. To create the predictive model, I decided to use forward selection, which cycles through each variable and interaction to find the best model to test the data.

III. Results

```
##
## Call:
## lm(formula = medianHouseValue ~ medianIncome + housingMedianAge +
##     households + population + latitude + longitude + totalBedrooms +
##     totalRooms + housingMedianAge:households + housingMedianAge:population +
##     population:totalBedrooms + housingMedianAge:latitude + housingMedianAge:longitude +
##     medianIncome:households + population:totalRooms + households:population +
##     medianIncome:housingMedianAge + medianIncome:latitude + medianIncome:longitude +
##     latitude:totalRooms + latitude:totalBedrooms + latitude:longitude +
##     housingMedianAge:totalBedrooms + housingMedianAge:totalRooms +
##     medianIncome:totalRooms + medianIncome:totalBedrooms + totalBedrooms:totalRooms +
##     longitude:totalRooms + longitude:totalBedrooms + medianIncome:population +
##     households:latitude + population:latitude + households:longitude +
##     population:longitude, data = housing_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -525311  -39281   -9535   28011  722486
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -8.511e+05  8.314e+05  -1.024 0.305995
## medianIncome                  -8.169e+05  4.982e+04 -16.399  < 2e-16 ***
## housingMedianAge              -9.254e+04  5.558e+03 -16.648  < 2e-16 ***
## households                    -1.485e+03  4.228e+02  -3.512 0.000445 ***
## population                     4.005e+02  1.343e+02   2.982 0.002869 **
## latitude                       9.887e+04  2.243e+04   4.407 1.05e-05 ***
## longitude                      5.810e+02  7.240e+03   0.080 0.936044
## totalBedrooms                 -6.872e+05  1.379e+05  -4.982 6.37e-07 ***
```

```
## totalRooms                       7.859e+05  1.429e+05   5.501 3.83e-08 ***
## housingMedianAge:households      7.602e+00  2.840e-01  26.771  < 2e-16 ***
## housingMedianAge:population     -2.150e+00  1.023e-01 -21.008  < 2e-16 ***
## population:totalBedrooms         1.857e+01  2.230e+00   8.329  < 2e-16 ***
## housingMedianAge:latitude       -1.222e+03  6.412e+01 -19.052  < 2e-16 ***
## housingMedianAge:longitude      -1.139e+03  6.436e+01 -17.696  < 2e-16 ***
## medianIncome:households          1.893e+01  2.199e+00   8.609  < 2e-16 ***
## population:totalRooms           -1.008e+01  1.541e+00  -6.537 6.44e-11 ***
## households:population            1.761e-03  3.671e-04   4.796 1.63e-06 ***
## medianIncome:housingMedianAge    3.327e+01  3.353e+01   0.992 0.321000
## medianIncome:latitude           -1.170e+04  5.588e+02 -20.933  < 2e-16 ***
## medianIncome:longitude          -1.061e+04  5.691e+02 -18.645  < 2e-16 ***
## latitude:totalRooms              1.390e+04  1.255e+03  11.076  < 2e-16 ***
## latitude:totalBedrooms          -1.172e+04  1.315e+03  -8.907  < 2e-16 ***
## latitude:longitude               5.985e+02  1.845e+02   3.244 0.001180 **
## housingMedianAge:totalBedrooms   5.213e+02  1.280e+02   4.071 4.70e-05 ***
## housingMedianAge:totalRooms     -2.254e+02  1.247e+02  -1.808 0.070625 .
## medianIncome:totalRooms         -1.418e+03  4.088e+02  -3.468 0.000527 ***
## medianIncome:totalBedrooms       1.961e+03  5.192e+02   3.776 0.000160 ***
## totalBedrooms:totalRooms        -1.140e+02  4.126e+01  -2.764 0.005724 **
## longitude:totalRooms             1.071e+04  1.512e+03   7.079 1.51e-12 ***
## longitude:totalBedrooms         -9.188e+03  1.498e+03  -6.132 8.85e-10 ***
## medianIncome:population         -2.370e+00  7.557e-01  -3.136 0.001716 **
## households:latitude             -2.356e+01  4.701e+00  -5.012 5.45e-07 ***
## population:latitude              6.871e+00  1.575e+00   4.363 1.29e-05 ***
## households:longitude            -1.834e+01  4.822e+00  -3.804 0.000143 ***
## population:longitude             5.223e+00  1.547e+00   3.376 0.000738 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 65200 on 16478 degrees of freedom
## Multiple R-squared:  0.6824, Adjusted R-squared:  0.6817
## F-statistic:  1041 on 34 and 16478 DF,  p-value: < 2.2e-16
```

The resulting model has many features and interactions, with a relatively high R^2. Many of the variables produced include interactions with either longitude or latitude, which would be expected as location is an important determinate of home values.

Out-of-sample accuracy

```
## [1] 66090.23
```

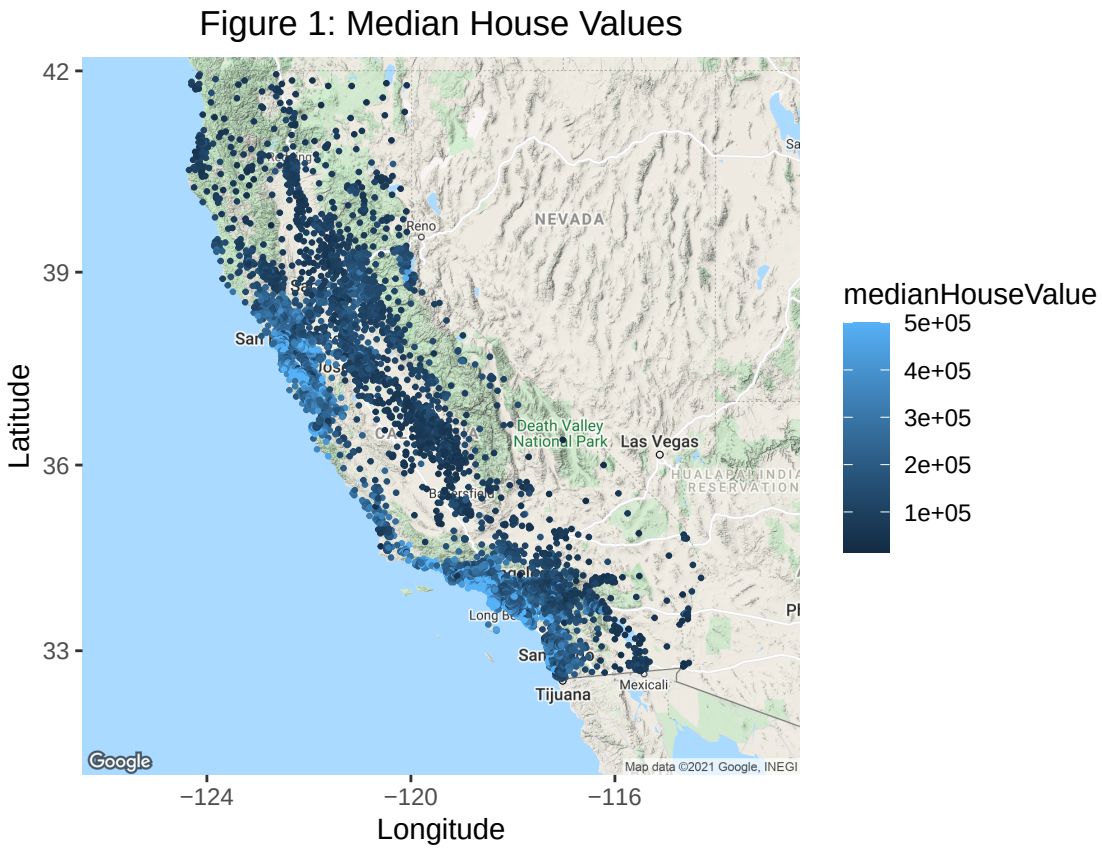


Figure 1 displays the original data set using the respective coordinates to plot the points on a map of the state of California. Upon observation, the highest median house values are along the coast, particularly in the San Francisco and Los Angeles regions, with several high median house values in the Lake Tahoe area. In contrast, the lowest median house values are in the Central Valley of the state.

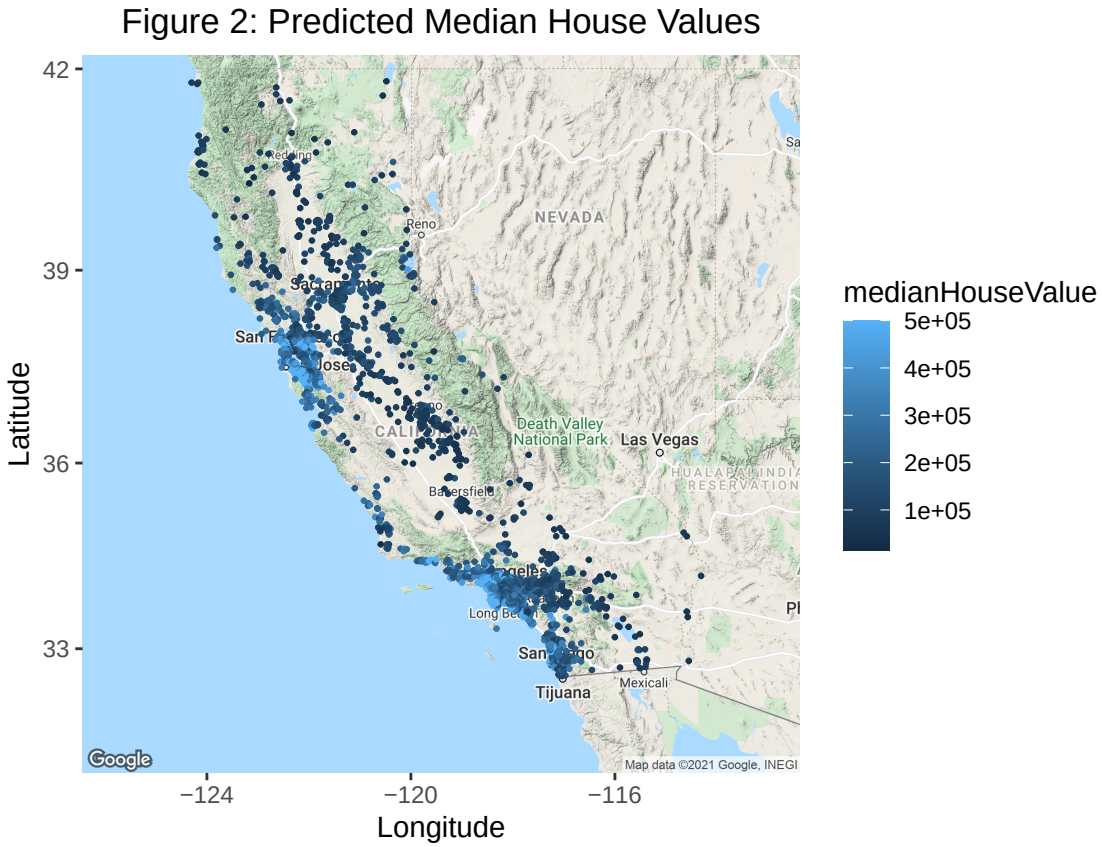


Figure 2 displays the predictive model of median house values using the original data set. As previously found in Figure 1, the data points display a clear indication of highest median house values along the California coast and several in the Lake Tahoe area, while the lowest median house values are in the Central Valley.

Figure 3: Residuals

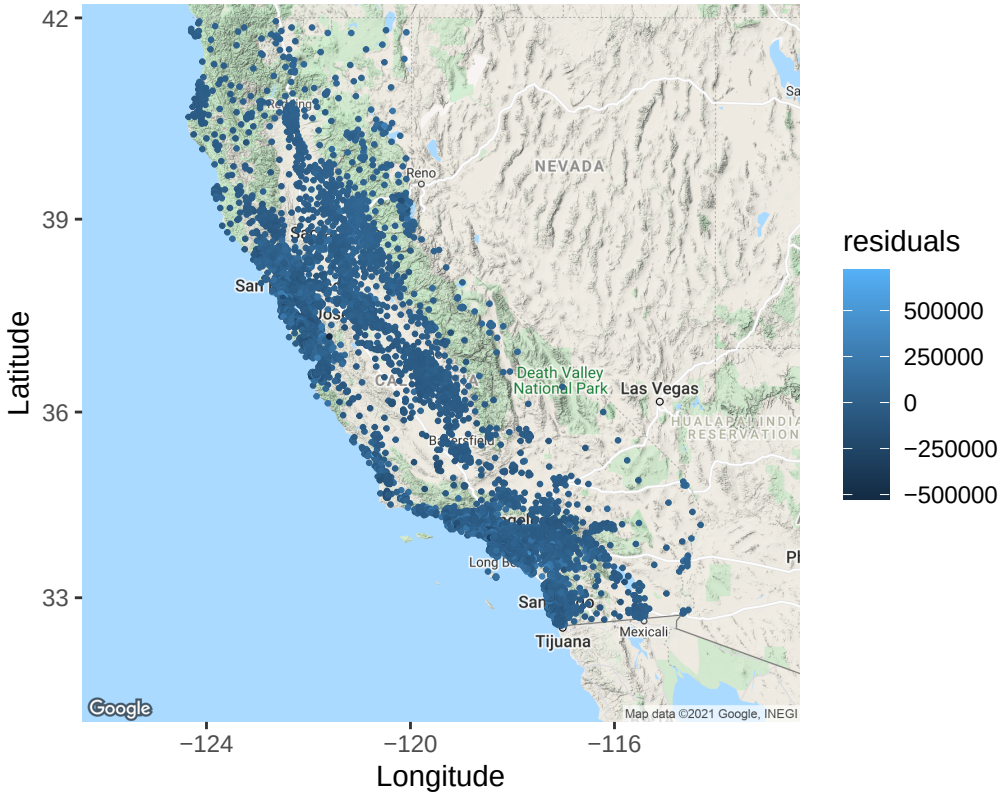


Figure 3 displays the residuals of the predictive model using the orginal data set. As expected, the map produced displays most data points centered around 0, with several outliers. This conveys that the created model was successful in determining the median house values.

IV. Conclusion

Using feature selection as a means to develop a model has been proven to be successful in predicting median house values across the state of California. Median house prices are highest along the coast and lowest in the Central Valley, which were expected and derived the original data set. The out-of-sample accuracy for the model is consistently close to 67000. Moreover, an obvious pattern to median house values has been detected and can be used for further analysis.