Cordero_Final_Project

Joaquin Cordero

2024-08-11

Having lived in California for most of my life now and trying to continue to live here for my adult life made me realize how expensive it truly is to live here. CA is notorious for being an expensive state to live in, although some may say CA residents also receive the highest income compared to their counterparts at other states, is it truly affordable to live here? From what I have seen, to find a place on your own is very difficult, so some would look for roommates in order to be able to afford housing. With the overall housing prices increasing across the US, wage has not been increasing relatively along with the rent/housing prices. Therefore, I wanted to see which cities would be the most affordable and safest place to live in.

First, the problem statement addressed is, with the rising concerns of affordable housing in the US I wanted to research which cities in CA would be the most affordable and safest place to live in. Although this will only address the state of CA, I believe it will be a good comparison for other people living in different states to see how they compare with CA residents. By showcasing the results from the results of the data, readers would have the information needed in order to see if they are able to live comfortably in the state of California. Showing which cities in the state has the lowest crime rate would also bring awareness to those that maybe want to move but is unsure of the area being safe. Also, providing a median income may give insight if they are able to maintain the same comfort similar to those that live in the same or surrounding area.

With all that, researching this topic and gathering data from multiple sources was the start of the analysis. The research conducted included 3 datasets, 2 from Kaggle and 1 from HDPulse. After realizing what information the datasets included, I decided to change the problem statement addressed from cities to

counties. Changing it to counties gave a better overall picture of the problem statement addressed, it also shortens the list and would make it easier to understand the overall results. An issue with the dataset from HDPulse came up when trying to load it into R, therefore some data cleaning, i.e. removing a couple header columns, were done outside of R. After loading all three datasets into R, counties and their respective data: median income, median house prices, and crime rate were taken out of each datasets. Cleaning it further by removing any duplicate entries based on counties from all the dataset to ensure we are able to keep the information for county observations. Finally combining all the cleaned datasets into 1 complete data set with all the necessary observations needed to begin the analysis portion. The data cleaning process below is the process conducted to start the analysis portion.

```
ca_house_data <- read.csv("California_Housing_CitiesAdded.csv")

ca_crime <- read.csv("rows.csv")

ca_income <- read.csv("HDPulse_data_export.csv") %>%
  rename('median_household_income' = Value..Dollars.) %>%
  slice(-c(1, 2))


my_data_house <- ca_house_data %>%
  select(Median_House_Value, City) %>%
  distinct(City, .keep_all = TRUE) %>%
  arrange(City) %>%
  rename('County' = City)


my_data_income <- ca_income %>%
  select(County, median_household_income) %>%
  mutate(median_household_income = median_household_income * 1000) %>%
  arrange(County) %>%
  mutate(County = str_remove(County, " County"))
```

```r
my_data_crime <- ca_crime %>%
  select(county_name, ratex1000) %>%
  slice(-c(1,2,3,4,5)) %>%
  distinct(county_name, .keep_all = TRUE) %>%
  rename('County' = county_name) %>%
  rename('crime_ratex1000' = ratex1000)


combined_data <- my_data_income %>%
  inner_join(my_data_house, by = "County") %>%
  inner_join(my_data_crime, by = "County")


affordableHousing <- combined_data %>%
  arrange(Median_House_Value)


safest_county <- combined_data %>%
  arrange(crime_ratex1000)


highest_income <- combined_data %>%
  arrange(median_household_income)
```

Since, the problem statement addressed is trying to show which counties would be the best and worst place to live in, showing the first and last 5 counties from the dataset would show its ranking status. To achieve this, arranging based on certain criteria, we are able to use the head and tail function to show the 5 best and worst counties in California. Below are the figures utilizing both functions and showcasing the results based on the criteria.

```r
head(affordableHousing, 5)
```

5 affordable counties

```
##     County median_household_income Median_House_Value crime_ratex1000
## 1    Modoc                  54000              69100            5.10
## 2   Sierra                  61000              93000            2.89
## 3   Colusa                  69000             123400            3.76
## 4    Glenn                  64000             125000            2.75
## 5 Siskiyou                  53000             128800            3.35
```

```r
tail(affordableHousing, 5)
```

5 least affordable counties

```
##           County median_household_income Median_House_Value crime_ratex1000
## 51     San Mateo                  149000             500001            3.11
## 52 Santa Barbara                   92000             500001            4.33
## 53   Santa Clara                  153000             500001            3.17
## 54    Santa Cruz                  104000             500001            4.30
## 55       Ventura                  102000             500001            2.77
```

```r
head(safest_county, 5)
```

5 most safe counties

```
##      County median_household_income Median_House_Value crime_ratex1000
## 1 Calaveras                  77000             133000            1.18
## 2    Plumas                  67000             130500            1.66
## 3  Tuolumne                  70000             217500            2.37
## 4    Lassen                  59000             200000            2.54
## 5    Placer                 109000             394000            2.55
```

```
tail(safest_county, 5)
```

5 least safe counties

```
##           County median_household_income Median_House_Value crime_ratex1000
## 51        Tehama                  59000             137900            7.34
## 52    Sacramento                  84000             448300            7.89
## 53       Alameda                 122000             500001            8.49
## 54 San Francisco                 136000             500001            8.85
## 55   San Joaquin                  82000             475000            9.07
```

```
head(highest_income, 5)
```

5 lowest median income counties

```
##     County median_household_income Median_House_Value crime_ratex1000
## 1 Imperial                  53000             159900            4.22
## 2 Siskiyou                  53000             128800            3.35
## 3    Modoc                  54000              69100            5.10
```

```
## 4     Lake                    56000              200500            5.01
## 5 Humboldt                    57000              165600            3.15
```

```
tail(highest_income, 5)
```

5 highest median income counties

```
##           County median_household_income Median_House_Value crime_ratex1000
## 51       Alameda                  122000             500001            8.49
## 52 San Francisco                  136000             500001            8.85
## 53         Marin                  142000             500001            2.56
## 54     San Mateo                  149000             500001            3.11
## 55   Santa Clara                  153000             500001            3.17
```

A linear model was attempted during the analysis part of the research as well. Although I would suggest a different approach, clustering model would probably serve useful as well when it comes to seeing the overall picture. By being able to use a cluster model all observation from the dataset would be accounted for and would group unlabeled examples to those that are similar with the 5 best/worst counties.

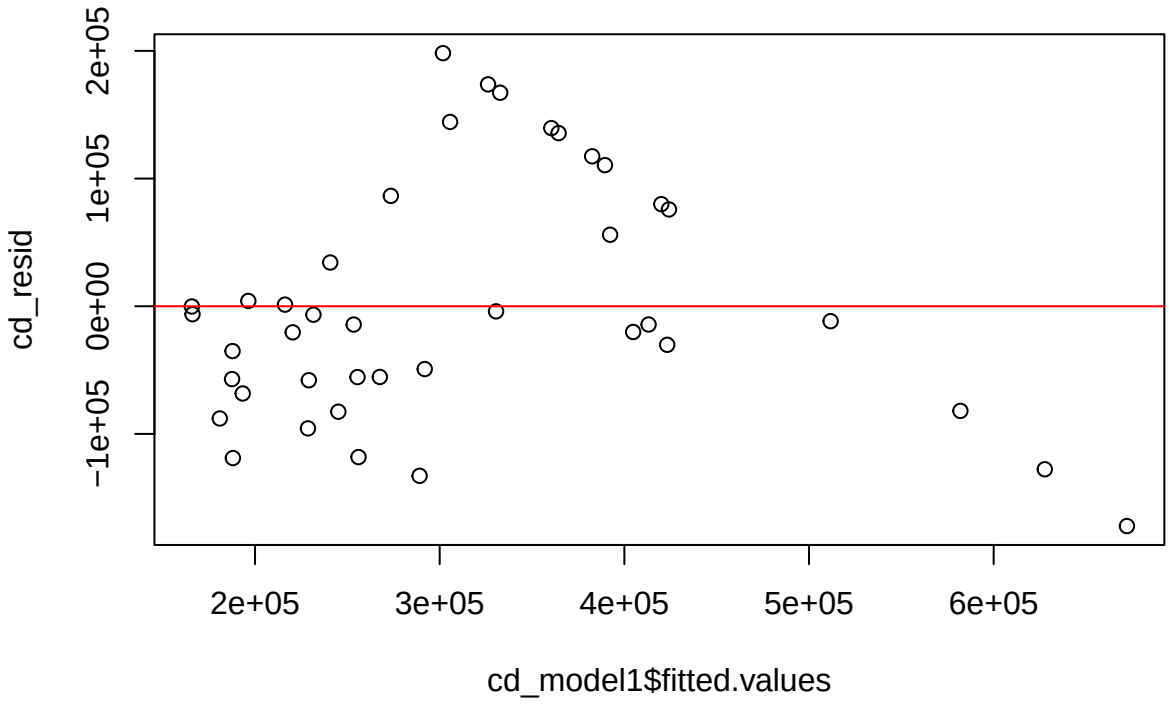
```
cd_split <- sample.split(combined_data, SplitRatio = 0.8)


cd_train <- subset(combined_data, cd_split == "TRUE")
cd_test <- subset(combined_data, cd_split == "FALSE")


cd_model1 <- lm(Median_House_Value ~ median_household_income
                + crime_ratex1000, data = cd_train)


cd_resid <- residuals(cd_model1)
```

```
plot(cd_model1$fitted.values, cd_resid) + abline(h=0, col = "red")
```




```
## integer(0)
```

The linear model above does not do a great example at representing the relationship properly since the model is not evenly distributed on the horizontal line at 0. The most interesting takeaway from the analysis was the expectation that safer counties would be the most expensive. The assumption was that higher crime rates would lower the overall housing prices in the county and that lower crime rates would raise overall housing prices. However, the tables above shows that even most expensive counties still has a higher crime rate and vice versa, some of the least expensive counties have a lower crime rate. Although, the median income does not really reflect the overall affordability or crime rate of the county, it is still good insight to see how it compares the median housing prices and its respective crime rates.

Being able to conduct this research, we are able to expand further on the overall problem statement. By

providing a list of the best and worst counties, we can go further into what other qualities makes them the best and worst. Other factors play into making housing more expensive as well as crime rates in the area. A view to the median income per county might not be a good representation to the overall picture of the counties, but it may play a role on the county development and why might some counties be placed a little higher or lower in the list. There are many factors that determines a countie's overall qualifications to be ranked the best and worst, I believe expanding on those topics and fiding what works or what does not work might help develop the overall status of the counties in California.

Lastly, the datasets used in the research does not include the population per county, having the population per county might have been usefull in the overall picture. The density of the population could show a better understanding at why the overall housing prices and crime rate would be higher or lower. Using a better model to be able to use all observations would also be beneficial as grouping the best and worst counties could prove useful, since discovering what makes them the best or worst could help bring light into the many problems the worst counties are facing to help bring them close to the status of the best ones. Using a different cleaning method of each data could also show different results from the same datasets, although the cleaning method here may not be the best method used, however it was the most ideal method conducted during the preparation before the analysis.

Overall, the main problem addressed was to find the most affordable and safest counties to live in California, by providing a list based on each criteria we are able to see the results. By showing the counties ranked the best and the worst readers are able to compare what makes the California counties the best and how it compares to the counties they would want to live in. Showing the median income to these counties also proves useful as one can determine if similar counties could be affordable based on their personal income. All things considered, California is notorious for being an expensive place to live in, many factors play into why it makes this place so expensive, but for now we are able to see where we could possibly live comfortably with the hopes that the area we chose would be safe enough for our everyday life.