Final Report - DataHacks 2020

Joshua Hong

Annie Tong

Luke Sztajnkrycer

2/9/2020

Table of Contents

1. Cleaning

I. Row Cleaning

To maximize the amount of data retained after row cleaning, the data set was selectively cleaned after relevant columns were selected from the available data. This allowed rows that had data in relevant columns but null values in other columns to be retained. Null values in the data set were represented as "...", so these values were replaced with null and then dropped from the data set. Other symbols that were removed from the dataset were "$" and "," which both caused issues when changing data types. While these symbols accounted for most of the row cleaning, there were some fringe cases, including empty strings and hyphens in the data. Numerical values in the data set were also converted from objects into floats to allow visualization and analysis.

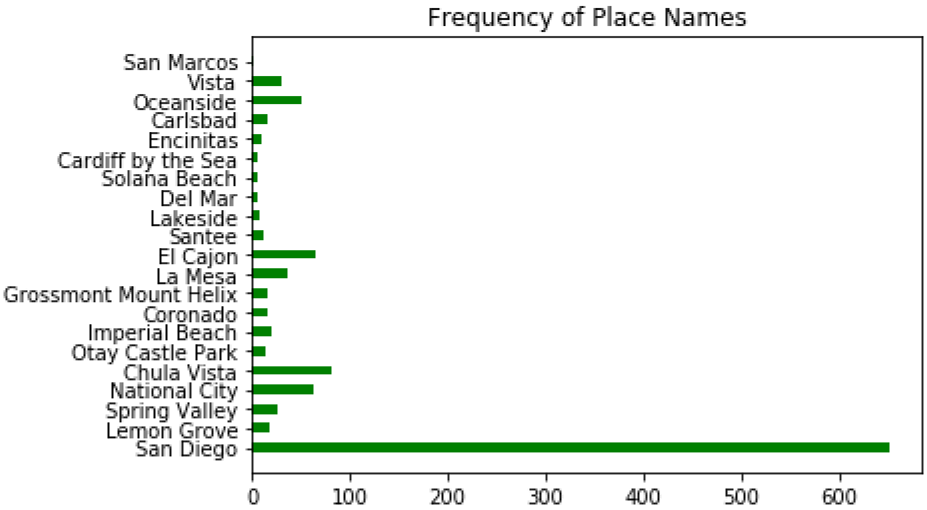
II. Column Cleaning

Column manipulation varied from model to model, as most models required different types of data. However, there were some columns that contained data that was widely used in many of the models, including the total number of persons and the average price of housing. From the raw data set, we constructed a new data set that removed many of the columns that were not the subject of our examination. Simultaneously, we combined similar data into possibly useful columns, including "Total male persons" and "Total Female persons". For example, the total male persons and the total female persons in census tract 1, block 1 is 405 and 496, respectively.

Census Tract	Block	Total Male	Total Female	Total Persons
1	1	405	496	901

2. Visualization

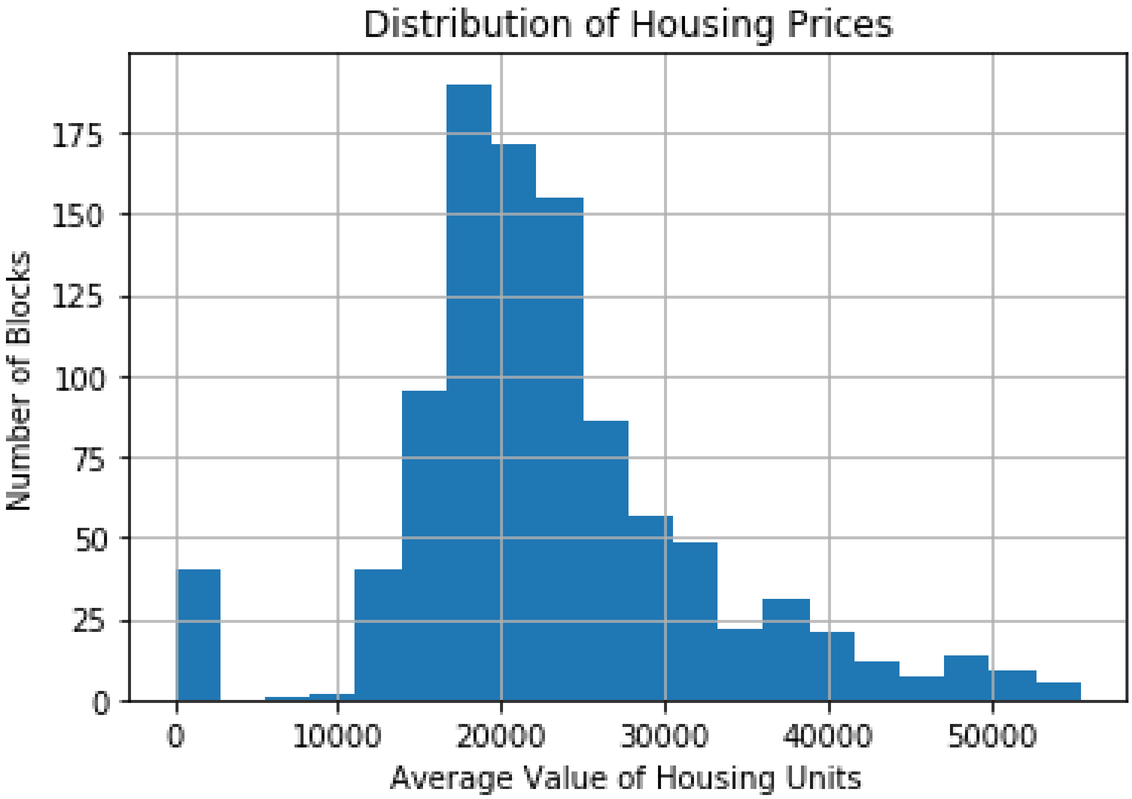
I. Block Sizes



```
{'San Diego': 652, nan: 112, 'Lemon Grove': 19, 'Spring Valley': 26, 'National City': 62, 'Chula Vista': 81, 'Otay Castle Park': 14, 'Imperial Beach': 20, 'Coronado': 16, 'Grossmont Mount Helix': 16, 'La Mesa': 37, 'El Cajon': 66, 'Santee': 12, 'Lakeside': 8, 'Del Mar': 5, 'Solana Beach': 5, 'Cardiff by the Sea': 5, 'Encinitas': 10, 'Carlsbad': 16, 'Oceanside': 50, 'Vista': 31, 'San Marcos': 1}
```

The most common place name is San Diego, followed by the second most common place name Chula Vista.

II. Distribution of Pricing of Houses

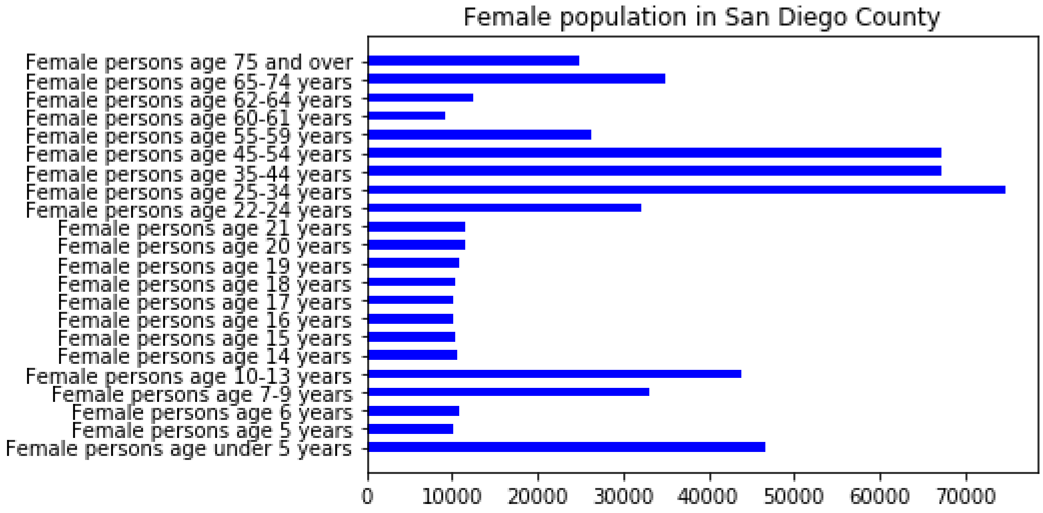


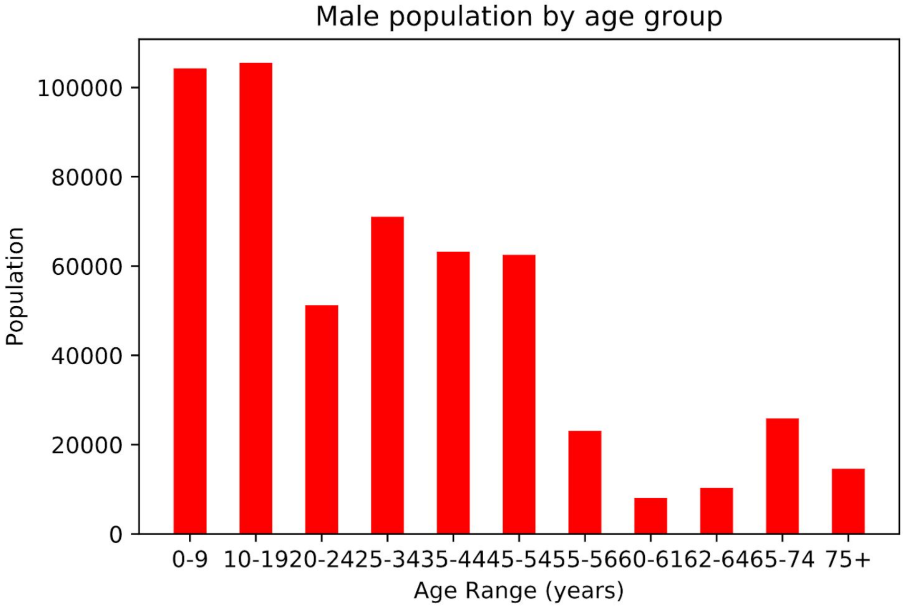
The average housing price in San Diego is roughly $23,030. This value was calculated by taking the average value of housing units in each block and multiplying by the total number of houses in the block. The resulting sum of total values was then divided by the total number of houses in San Diego.

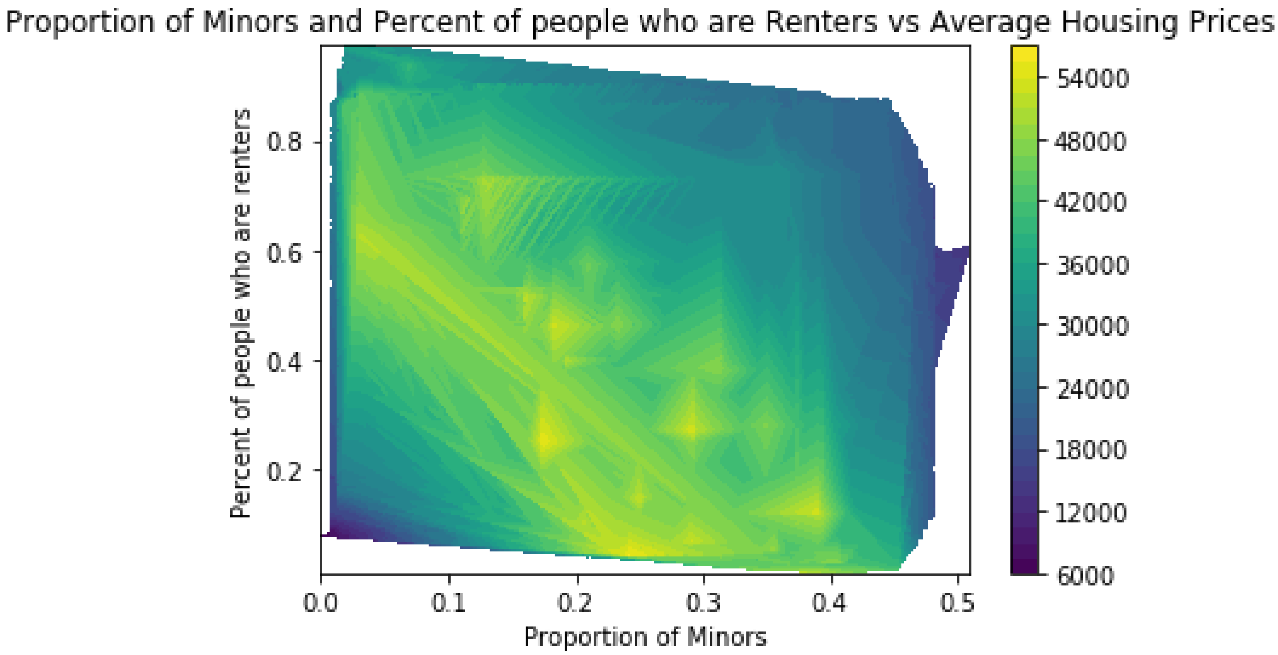
Another notable datapoint is the block that has the highest average value of housing units. After analyzing the data, we find that Census Tract 83.03, block group 1 has the highest average housing price at $55,508 with 281 houses.

Census Tract	Block	Average Price for Housing Units	Number of Housing Units
83.03	1	55508	281

III. Plotting of Various Categories





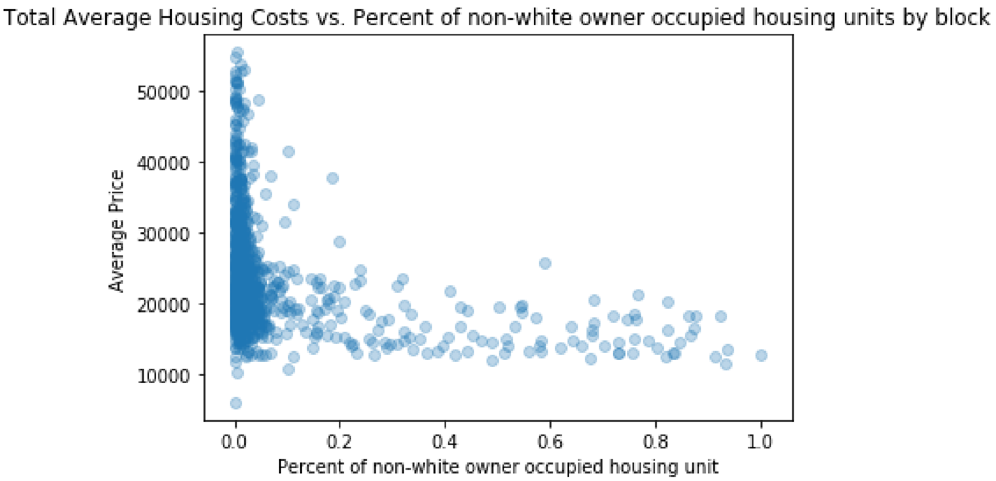


The graph above is a contour plot that graphs the proportion of minors and the percent of people who are renters against average housing prices. While the contours of the plot are not very smooth, the higher average housing prices do seem to fall on a rough line. This could serve as a topic of research for further analysis.

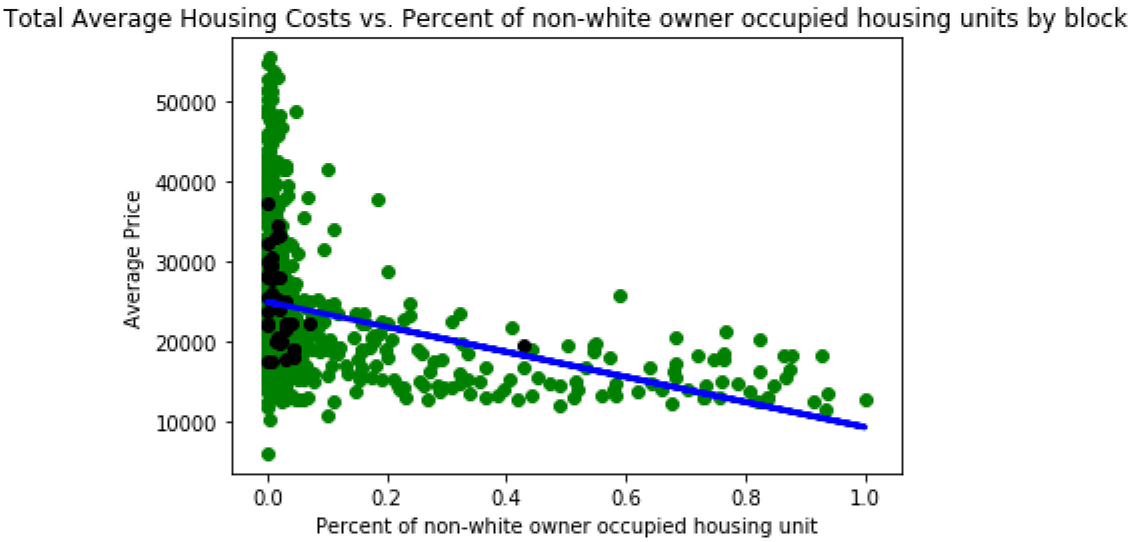
3. Machine Learning (Linear Regression)

I. Living Conditions of Houses

We also wanted to see if there was a correlation between average housing costs and the presence of colored/non-white individuals by block. Due to the history of social restriction placed on non-caucasian individuals in the U.S., and the stereotypically lower socioeconomic status that results of this oppression, we were curious to see if a greater non-white population was correlated with less expensive housing, on average. In order to measure possible trends, we compared the total average housing costs to the percent of non-white home owners for each block. The percent of non-white home owners was calculated by determining the quotient of white owners occupying housing units divided by the total of owners occupying housing units, and subtracting this value from one. The results:

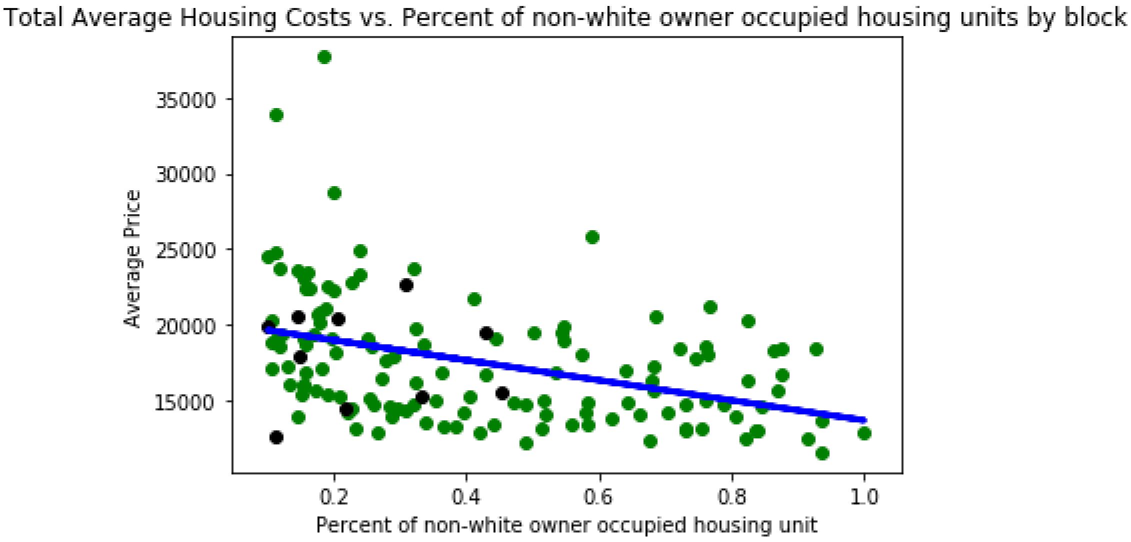


This seemed promising. As one can see, in blocks with low densities of non-white individuals (and conversely high densities of white individuals), there is a wide range of price variance: avergace house prices range from under $10,000 to over $50,000. This price range seems to become more concentrated as the concentration of non-white individuals surpasses ~10%, now only stretching between $15,000 and $25,000. So, we wanted to try to find a linear correlation for this comparison.



Coefficient of determination with respect to black points: 0.07, Coefficient of determination with respect to green points: 0.10

Unfortunately, we could not find a very strong linear regression model to represent this data. We also tried trimming the dataset to not include blocks with 10% or less non-white individual makeup. This allowed us to only examine blocks with non-white densities of 10% or higher.



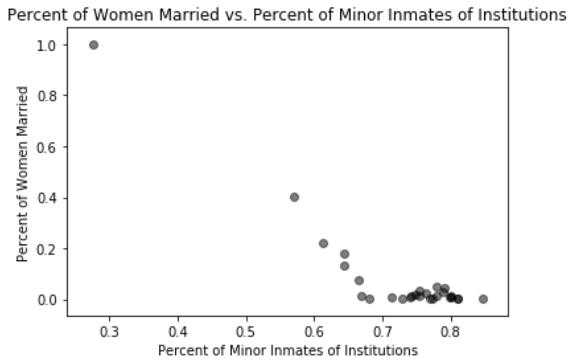
Coefficient of determination with respect to black points: 0.13, Coefficient of determination with respect to green points: 0.17

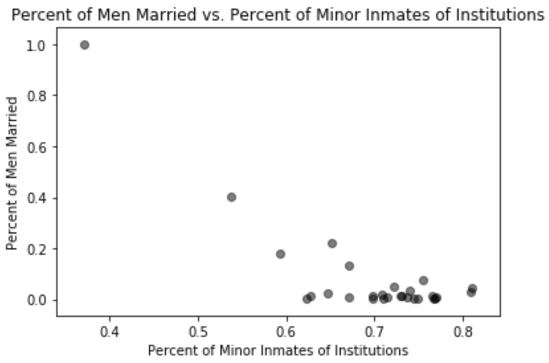
This did not yield a very strong r-squared value either. The data has a weak negative correlation but is still relatively spread out. These graphs show that it is not fair to predict a block's average house price by % non-white race makeup (due to the low coefficients of correlation), however our very first graph showing the trend over all useable data shows that different races have different access to housing, or that race makeup does effect housing prices. We can see that blocks possessing ~10% non-white owner occupied housing units or higher are virtually excluded from housing with average prices above $25,000, while there is a much wider price range for blocks below this 10% threshold. Beyond this, linear regressions will not allow us to infer anything significant regarding how a continually increasing non-white race makeup affects average housing prices.

II. Percent Married vs Percent of Minors that are Inmates of Institutions

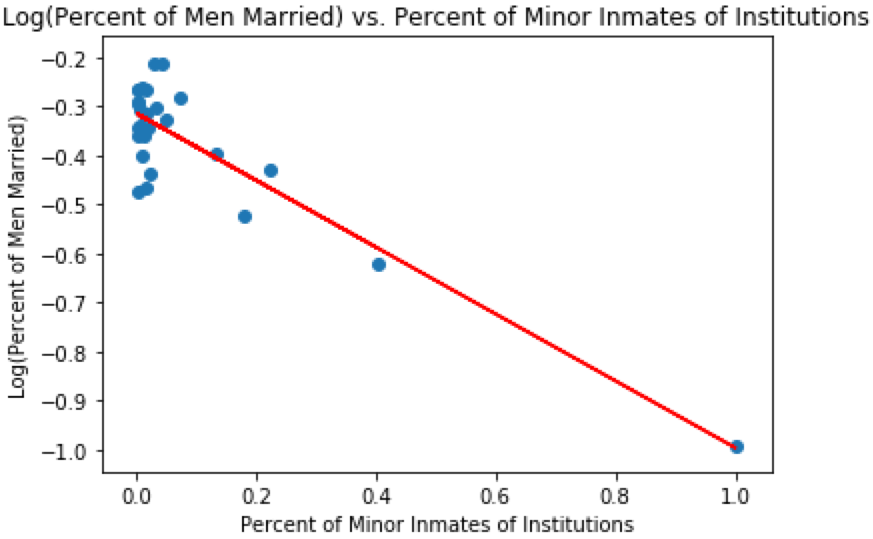
In another area, we studied the relationship between the percent of minors who have been incarcerated versus the percent of men and women married in each block. Our hypothesis was that the more unmarried parents there were in an area, the more likely that area would have minors who are inmates of institutions. Several studies before this have argued that troubled teens tend to come from families in which at least one parent figure was missing, and we wanted to test if this would hold true for the San Diego County communities at least in the 1970s.

With the few blocks that chose to release information on the number of inmates of institutions, we graphed those 28 block data points according to the relationship we were trying to test. First we graphed percent of women married versus percent of minors incarcerated, then we graphed percent of men married versus percent of minors incarcerated, as shown below.

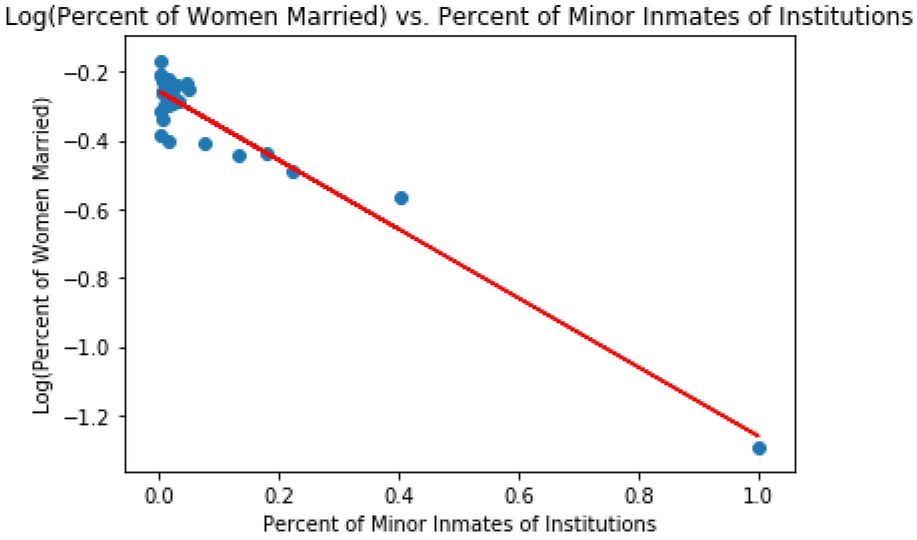




There is at least some indication of a trend in the graphs above. The graphs seem to indicate a logarithmic relationship to some extent, so we tested that out in the next graphs shown below, along with our linear regression model for said graphs, by taking the log of percent of men married and the log of percent of women married.



Coefficient of determination: 0.79

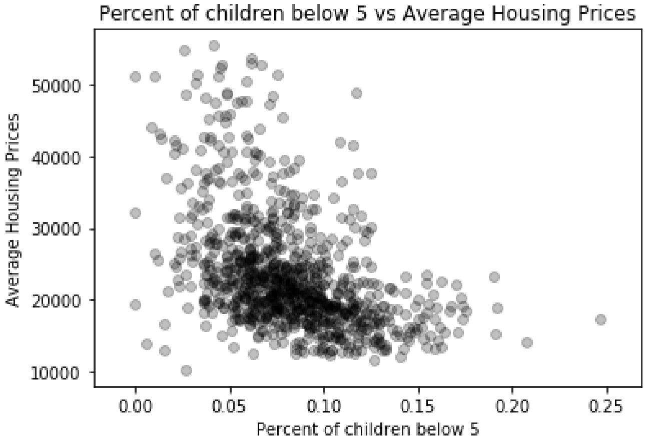


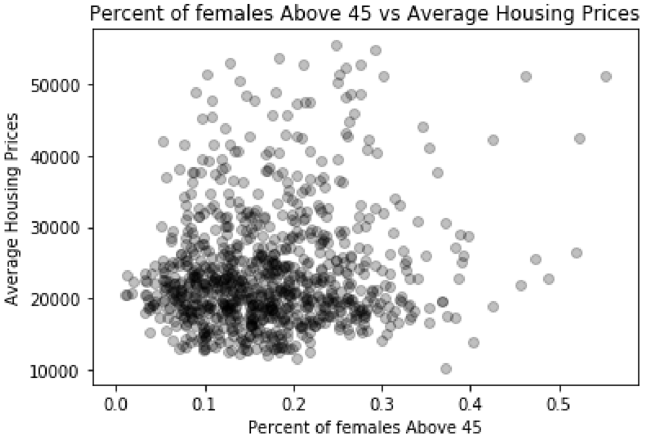
Coefficient of determination: 0.92

As seen above, a relatively strong trend can be seen between the two factors. There can be at least some correlation made between the percentage of parents married versus the percent of minors who are inmates of institutions. Much of the studies done on the relationship between these two factors were meant to disprove the long-standing belief that race, not family conditions, affected the likelihood of minors to commit crimes. If we were to further our research, we would also look at the relationship between percent of minors that are inmates of institutions and percent of total people who are black, if enough data is made available to properly study this relationship, in order to possibly refute arguments that one's race determined one's likelihood of committing crimes.

III. What is the Relationship between Age Distribution and Rent Statistics?

A third area we looked into was the relationship between age distributions in a block and the percent of people who live in rented housing. Initially, we tried to find a relationship between very young children and the average price of housing units, but the resulting graph showed highly variable data. We also tried the same relationship except with persons above age of 45,

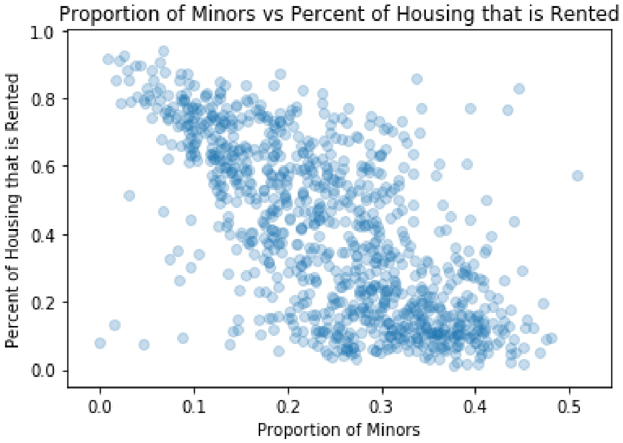

Percent of children below 5 vs Average Housing Prices

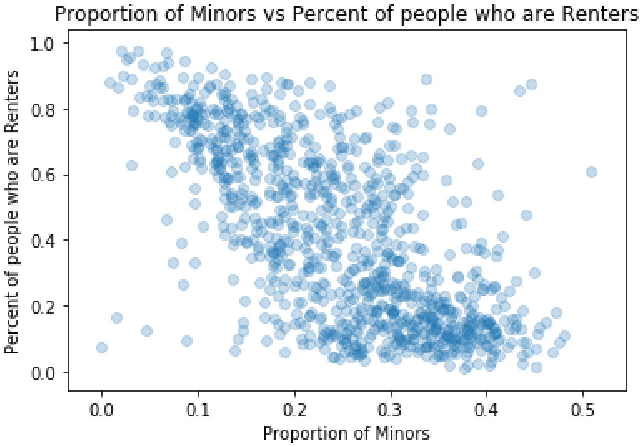

Percent of females Above 45 vs Average Housing Prices

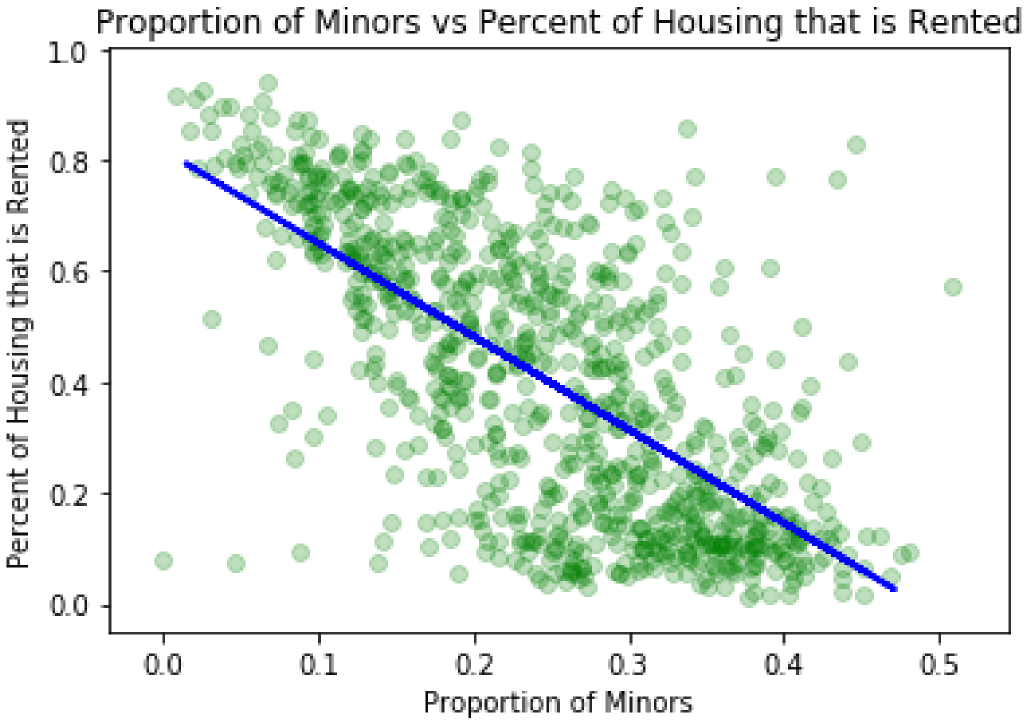
which has similar results.

While the data was highly variable, there did seem to be some trend between age distributions and housing, so we decided to instead see whether age distribution was correlated with the percentages of people who rented. Using the data from the original data set that described the number of persons under 18 in a family, we defined a minor to be a person under 18 living in a household. After cleaning the data and calculating the proportion of minors to the total population for the block, we got a graph that had less variation. The resulting graph showed a similar trend for the percent of housing in an area being rented and the percent of people who are renters.

There seems to be some correlation between the variables in the figures above, so using sci-kit we created a linear regression model to see how strong the correlation was.

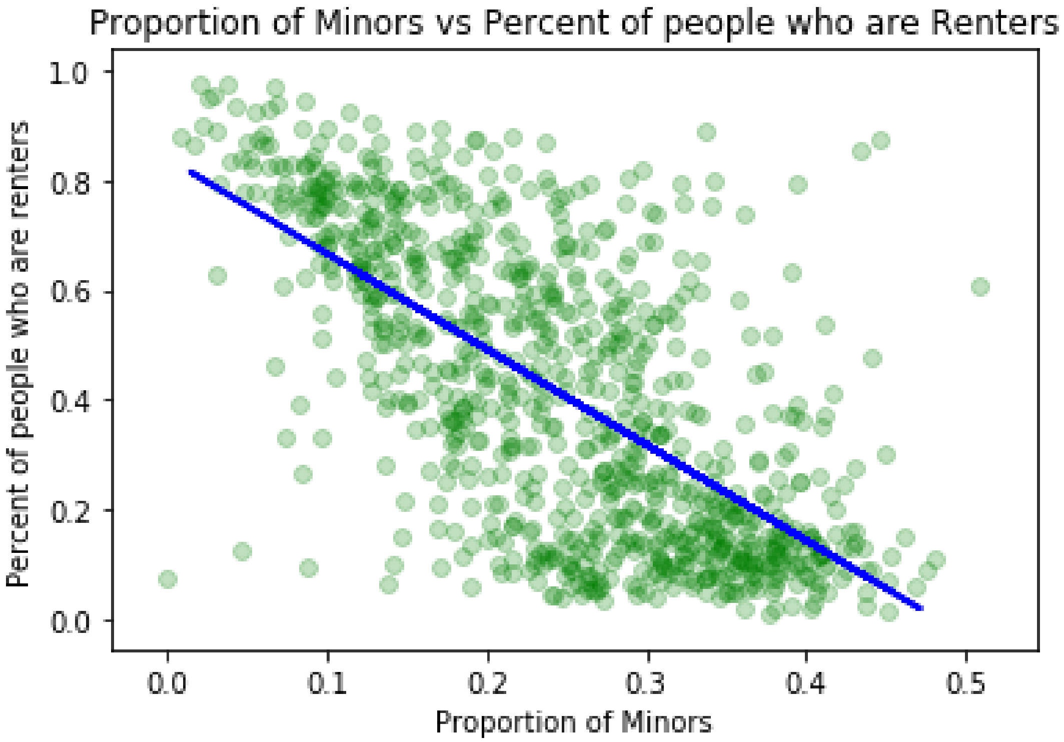

Proportion of Minors vs Percent of Housing that is Rented


Proportion of Minors vs Percent of people who are Renters



Mean squared error: 0.03, Coefficient of determination: 0.46

From the figure, we can see that the linear regression line models the data relatively well. The coefficient of determination for the model is 0.46, which indicates that 46% of the variation in the percent of housing that is rented can be explained by the variation in the proportion of minors.



Mean squared error: 0.04, Coefficient of determination: 0.45

Similar to the model above, there seems to be some correlation between the proportion of minors and the percent of people who are renters. The coefficient of determination for the model is 0.46, which indicates that 45% of the variation in the percent of housing that is rented can be explained by the variation in the proportion of minors. These two models have similar coefficients of determination and mean squared errors, indicating that they represent the data equally well.

From the models, we can see that there appears to be some correlation between age distribution and rent statistics. It’s possible that families with minors are more inclined to stay in one area for an extended amount of time, which could possibly allow for children to have a more consistent environment while growing up. This observation could account for how blocks with a higher percentage of children tend to have less rented houses and people living in rented housing, as families with children prefer to buy their own houses in contrast to living in a rented house. Further analysis could possibly be done on individual families in San Diego where information such as rent information, age of children, and years spent living in the same house. This information could be used to see further trends between age distributions and relocation tendencies of families.

4. Links

- https://github.com/annieteaaa/DataHacks2020 (Under Dataset)