Jared Yu

Data Visualization

Project 3

Note: The code for the assignment is in the end as an appendix.

The Data (same as Project 2)

The dataset chosen for this project is the California Housing Prices dataset located on the website 'Kaggle'. It was posted by the user 'Cam Nugent,' a postdoctoral bioinformatics researcher from the University of Guelph in Ontario, Canada. According to the user's Kaggle profile, they are a user that is skilled in curating datasets, having achieved the rank of Expert in the category. The dataset is described as the median house prices for California districts derived from the 1990 census. On the page that the dataset is posted on, there are 188 votes in favor of the dataset, indicating that many users have benefited from practicing their data science skills on this data. There are also 91 kernels that show how different users have utilized the data and presented their data analysis to others. As a California native, this dataset intrigues me and is different from a previous dataset that I've investigated based off comparatively recent home prices for San Francisco Bay Area real estate.

The description of the data on the page that it's hosted on indicates that the dataset comes from Aurelien Geron's recent book, 'Hands-On Machine learning with Scikit-Learn and TensorFlow.' It contains the following ten variables: `longitude`, `latitude`, `housingmedianage`, `total_rooms`, `total_bedrooms`, `population`, `households`, `median_income`, `medianhousevalue`, `ocean_proximity`. The dataset was first featured in a paper published in 'Statistics & Probability Letters (1997)'. Also, according to the description of the data, the variables for the dataset are self-explanatory. However, in the file description, there are also further details that more fully explain each of the variables. For the purpose of this third project, additional variables were created through feature engineering. The new variables include: `class`, `city`, and `sfla`. The variable `class` refers to the lower, middle, and upper classes. Records in the dataset have their `class` determined by their median income level. The variable `city` is based off the `longitude` and `latitude` coordinates provided in each record. Lastly, the `sfla` variable determines whether a record exists in the San Francisco Bay Area, Greater Los Angeles Area, or other. More details for these features are explained later.

There's a total of 20,640 records in the dataset. The records themselves refer to houses found in California districts with summary statistics based on information from the 1990 census data. The dataset however is not cleaned, as indicated by 207 records that contain `NA`'s for values. After the rows with `NA`'s are removed, there are only 20,433 records remaining. Furthermore, during the feature engineering process some records had values of `NA` and so they were removed also, leaving 20,179 total records.

The Variables

Below is a table of all the different variables mentioned previously along with the variables that were created through feature engineering. The table provides some brief descriptive statistics along with explaining the category that each variable belongs to (e.g., quantitative / ratio). Towards the end there are a few categorical variables where descriptive statistics such as mean, median, and mode are not possible. Other techniques will be used later to examine these variables more closely.

Variable	Category	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum	Std. Deviation	Mode
longitude	Quantitative / Interval	−124.35	−121.8	−118.49	−119.571	−118.01	−114.31	2.003578	−118.31
latitude	Quantitative / Interval	32.54	33.93	34.26	35.63322	37.72	41.95	2.136348	34.06
housing_median_age	Quantitative / Ratio	1	18	29	28.63309	37	52	12.59181	52
total_rooms	Quantitative / Ratio	2	1450	2127	2636.504	3143	39320	2185.27	1527
total_bedrooms	Quantitative / Ratio	1	296	435	537.8706	647	6445	421.3851	280
population	Quantitative / Ratio	3	787	1166	1424.947	1722	35682	1133.208	891
households	Quantitative / Ratio	1	280	409	499.4335	604	6082	382.2992	306
median_income	Quantitative / Ratio	0.4999	2.5637	3.5365	3.871162	4.744	15.0001	1.899291	3.125
median_house_value	Quantitative / Ratio	14999	119500	179700	206864.4	264700	500001	115435.7	500001
ocean_proximity	Qualitative / Nominal	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
class	Qualitative / Ordinal	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
city	Qualitative / Nominal	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
sfla	Qualitative / Nominal	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

1. The first variable is `longitude`, which is defined as, “A measure of how far west a house is; a higher value is farther west.” It is a quantitative variable that is of type interval. The reason is that zero values for this variable don’t indicate that there’s no longitude. Below in Figure 1 is a histogram and boxplot of the variable. The histogram and boxplot both indicate that many of the records in the dataset are located have a latitude of around -118. Using the `median()` function, this area is estimated to be around -118.49. The calculated mode also indicates that the peak is around this number of -118.31. There also appears to be a second peak in the distribution, as the histogram seems to be bimodal in its shape. The alternate peak occurs around -122.

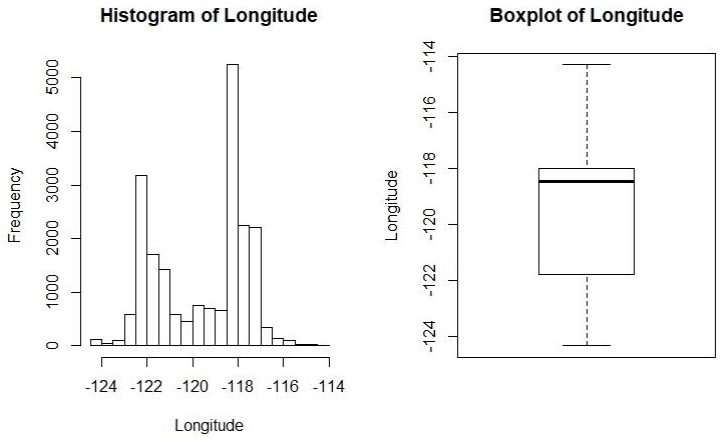

Figure 1

Furthermore, below is a table of the basic descriptive statistics for the variable. They reflect the data represented in the histogram and boxplot.

Minimum	1st Quantile	Median	Mean	3rd Quantile	Max	Std. Deviation	Mode
-124.35	-121.8	-118.49	-119.571	-118.01	-114.31	2.003578	-118.31

2. The second variable is `latitude`, which is defined as, "A measure of how far north a house is; a higher value is farther north." It is a quantitative variable that is of type interval. The reason is that zero values for this variable don't indicate that there's no latitude. Below in Figure 2 is a histogram and boxplot of the variable. Like the previous variable `longitude`, there is also another bimodal distribution evident in the histogram. Here, the data seems focused around 34 and alternately around 38. Using the function `median()`, the largest portion of the data seems concentrated around 34.26.

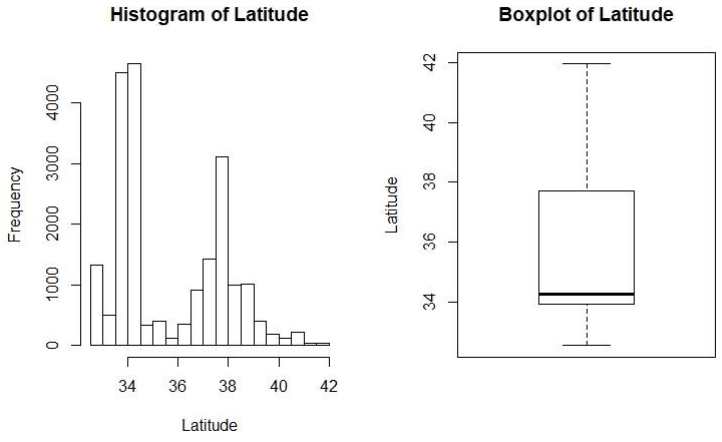

Figure 2

Furthermore, below is a table of the basic descriptive statistics for the variable. They reflect the data represented in the histogram and boxplot.

Minimum	1st Quantile	Median	Mean	3rd Quantile	Max	Std. Deviation	Mode
32.54	33.93	34.26	35.63322	37.72	41.95	2.136348	34.06

3. The third variable is `housing_median_age`, which is defined as, "Median age of a house within a block; a lower number is a newer building." This variable indicates the median age of houses within the block that a record is located. This is a quantitative variable that is a ratio. This variable is a ratio since the value zero is meaningful in that it would indicate that the house has no age or is brand new. However, it's difficult to imagine that a house could have no age in the dataset unless the entire set of houses within a block were built at the same time as when the dataset was gathered.

Using the `min()` and `max()` functions, the dataset does however show that there are records with an age of 1 and up to a maximum age of 52. Below in Figure 3 are a histogram and boxplot of the data. The distribution of the histogram seems roughly unimodal with peaks around 20 and 30 for the median age. The boxplot shows that the median is around 30, which is confirmed to be 29 using the function `median()` in RStudio. It doesn't look entirely obvious, but the data is possibly skewed towards the left, indicating that the records are mostly homes that are above the median age, with fewer newer houses. It's interesting however to note that the maximum and the mode are both 52, indicating possibly that this is the cap for this value in the dataset. Therefore, any home at or above 52 for this variable would be assigned this value.

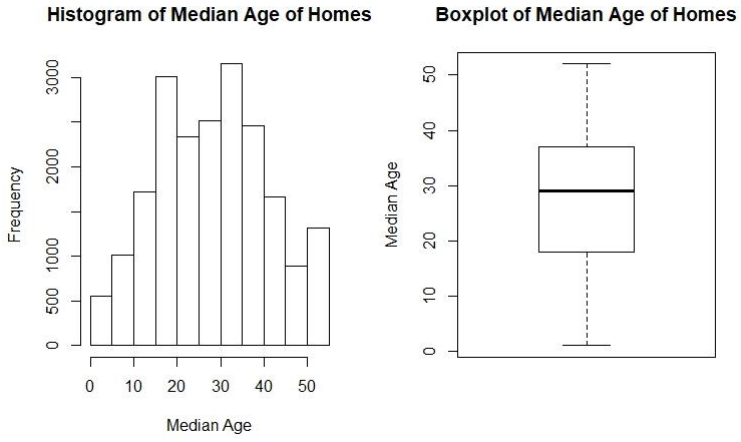

Figure 3

Furthermore, below is a table of the basic descriptive statistics for the variable. They reflect the data represented in the histogram and boxplot.

Minimum	1st Quantile	Median	Mean	3rd Quantile	Max	Std. Deviation	Mode
1	18	29	28.63309	37	52	12.59181	52

4. The fourth variable is `total_rooms`, which is defined as, "Total number of rooms within a block." This variable seems to indicate that each record exists within a defined block and each block has a discrete number of rooms. It is a quantitative variable also of type ratio. Below in Figure 4 are a histogram and boxplot of the variable. The histogram and boxplot both show that there's a strong skew towards the right in the data, with most of the data concentrated closer to 0. Using R, the minimum, maximum, and medium number of rooms within a block is 2, 39,320, and 2,127. The data for this variable contains many outliers and based on the median value it seems most records are within blocks that have around 2,000 rooms. However, looking at the boxplot it's apparent that there are outliers showing significantly more densely populated areas where there are many more rooms per block. This result can make sense if it were imagined that those outliers are in the most heavily densely populated areas leading to there being many rooms per block. Further away, where records are in less densely packed areas there are blocks of only about 2,000 rooms. Based on the distribution of populations in California, this result could make sense. Yet, further analysis would be needed to fully understand the details.

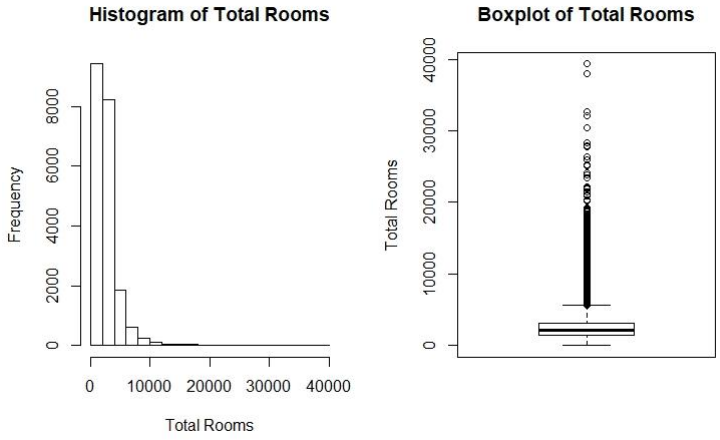

Figure 4

Furthermore, below is a table of the basic descriptive statistics for the variable. They reflect the data represented in the histogram and boxplot.

Minimum	1st Quantile	Median	Mean	3rd Quantile	Max	Std. Deviation	Mode
2	1450	2127	2636.504	3143	39320	2185.27	1527

5. The fifth variable is `total_bedrooms`, which is defined as, "Total number of bedrooms within a block." This variable seems to be like the previous variable, except it specifies the number of bedrooms per block. It is a quantitative variable also of type ratio. They don't seem to be too different based upon their definitions alone. Looking below at Figure 5 are the histogram and boxplot of the variable. The distributions for them also seem quite like the previous variable, with the difference being that they are scaled down. The minimum, maximum, and medium of the data is 1, 6,445, and 435. The skew is just like the previous variable and an explanation of the outliers would require deeper analysis.

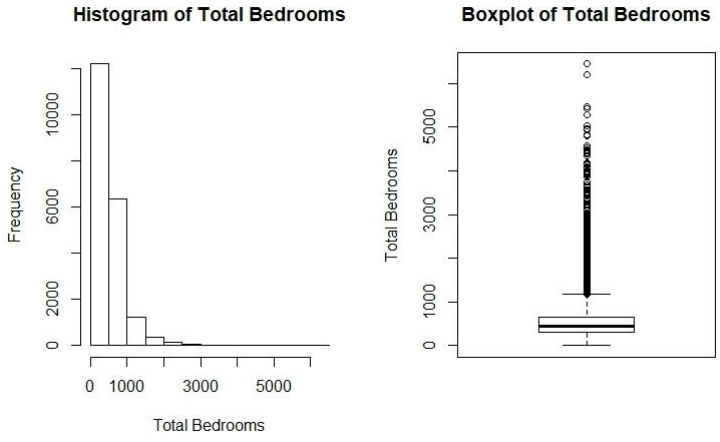

Figure 5

Furthermore, below is a table of the basic descriptive statistics for the variable. They reflect the data represented in the histogram and boxplot.

Minimum	1st Quantile	Median	Mean	3rd Quantile	Max	Std. Deviation	Mode
1	296	435	537.8706	647	6445	421.3851	280

6. The sixth variable is `population`, which is defined as, "Total number of people residing within a block." This variable indicates how many people live within a record's block. It is a quantitative variable of ratio type. However, it may not be entirely logical to think that a record of a home could have a zero value since that would indicate nobody living there. Yet, since this is census data, it would depend on how those statistics are surveyed from the population.

Below in Figure 6 are the histogram and boxplot of the variable. The distribution of the unimodal histogram is skewed far to the right and the boxplot shows that there are many outliers including two extreme outliers. The minimum, maximum, and median values are 3, 35,682, and 1,166. Although there are some areas with a highly dense number of residents per block, the majority seem to have only closer to around 1,000 residents. An interesting question would be to analyze the two extreme outliers evident in the boxplot, but this is a task that requires deeper analysis.

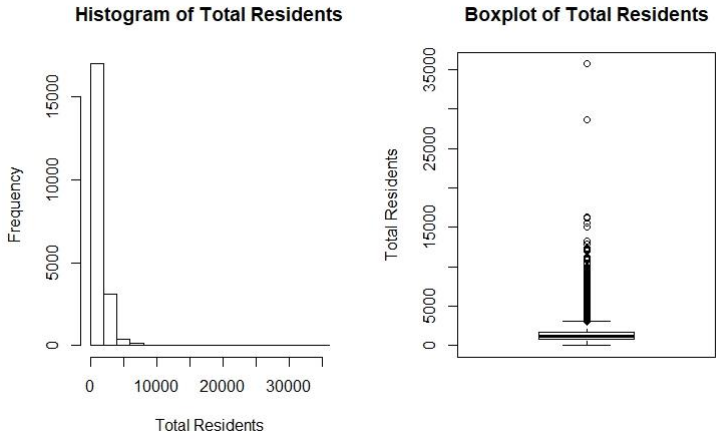

Figure 6

Furthermore, below is a table of the basic descriptive statistics for the variable. They reflect the data represented in the histogram and boxplot.

Minimum	1st Quantile	Median	Mean	3rd Quantile	Max	Std. Deviation	Mode
3	787	1166	1424.947	1722	35682	1133.208	891

7. The seventh variable is `households`, which is defined as, "Total number of households, a group of people residing within a home unit, for a block." It indicates the number of households or groups of people living within a record's block. This variable is a quantitative variable that is of type ratio. However, it's also unlikely that this variable would be zero, indicating that there are no households within a record's block. The reason is that this would exclude the record itself, which ideally should be a household. The minimum, maximum, and median of the variable is 1, 6,082, and 409. Below in Figure 7 are the histogram and boxplot of the variable. The histogram is unimodal with a strong skew to the right. The boxplot shows that there are many outliers in the data that exist in the higher end of the distribution. The pattern seen in this variable is like the pattern seen in other variables discussed.

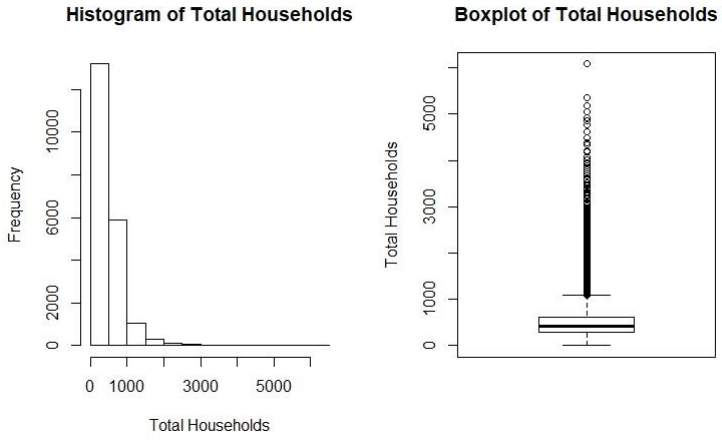

Figure 7

Furthermore, below is a table of the basic descriptive statistics for the variable. They reflect the data represented in the histogram and boxplot.

Minimum	1st Quantile	Median	Mean	3rd Quantile	Max	Std. Deviation	Mode
1	280	409	499.4335	604	6082	382.2992	306

8. The eighth variable is `median_income`, which is defined as, "Median income for households within a block of houses (measured in tens of thousands of US Dollars)." This variable gives the median income for homes within a record's block. This variable is a quantitative variable of type ratio. Although in theory the variable could be zero, it wouldn't make too much sense however since an income is required to pay for a home's mortgage. If a family couldn't afford to pay, then they would be removed and new tenants with income would move in. Also, it seems unlikely that an entire block of homes would have zero income.

The minimum, maximum, and medium values of the variable are 0.4999, 15.0001, and 3.5365. Below in Figure 8 are the histogram and boxplot for the variable. The histogram has a unimodal distribution with a skew towards the right, but the skew isn't as extreme as seen in previous variables. The boxplot shows a great deal of outliers at the higher end of the distribution.

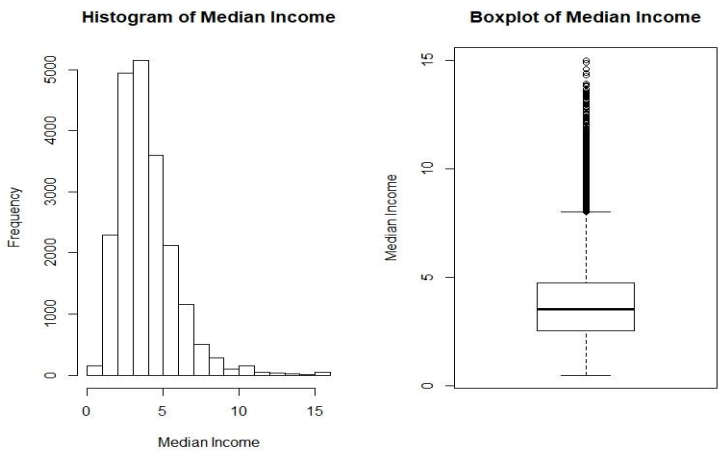

Figure 8

Furthermore, below is a table of the basic descriptive statistics for the variable. They reflect the data represented in the histogram and boxplot.

Minimum	1st Quantile	Median	Mean	3rd Quantile	Max	Std. Deviation	Mode
0.4999	2.5637	3.5365	3.871162	4.744	15.0001	1.899291	3.125

9. The ninth variable is `median_house_value`, which is defined as, "Median house value for households within a block (measured in US Dollars)." This variable indicates the median value of the homes within a block of a record. This is a quantitative variable of type ratio. In theory, the value of a home could be zero, but it's unlikely to ever happen and it's evident in the data that no such case exists. The distribution of the histogram appears unimodal with a skew towards the right. However, at the right end of the distribution is a small spike showing that there is a density of extreme values towards the higher end. The boxplot shows a similar picture with values centered in an area, and some outliers towards the end. The median of this data is 179,700 which represents the main peak in the distribution of the histogram. The outliers in the boxplot seem to be around 500,000. There's possibly some cap on this value, as the mode is also 500,001, identical to the maximum value. Then any areas with median values above that would all be lumped into this value. Using R, it says that there are 958 records that belong to this category. This is an interesting statistic that I feel is worthy of investigation.

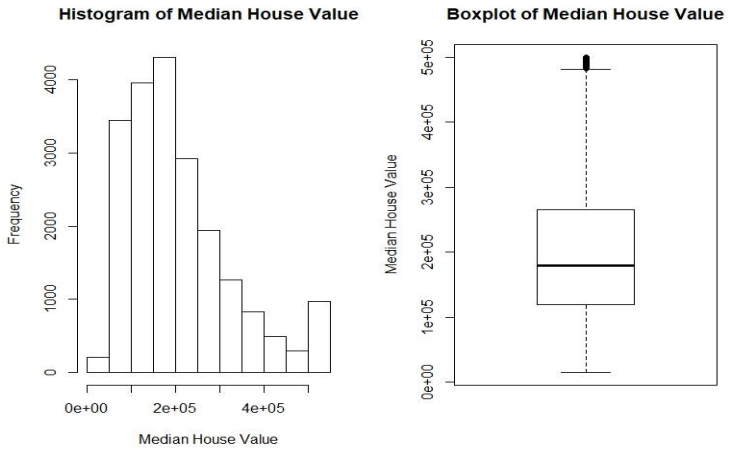

Figure 9

Furthermore, below is a table of the basic descriptive statistics for the variable. They reflect the data represented in the histogram and boxplot.

Minimum	1st Quantile	Median	Mean	3rd Quantile	Max	Std. Deviation	Mode
14999	119500	179700	206864.4	264700	500001	115435.7	500001

10. The tenth variable is `ocean_proxmity`, which is defined as, "Location of the house w.r.t ocean/sea." The variable indicates the distance that a record is from the ocean/sea. This is a qualitative variable or nominal variable. California is a state along the Pacific Ocean and so it makes sense that many homes would be located near the ocean. Also, there is the San Francisco Bay Area with a high density of human population. Therefore, it makes sense that the level 'NEAR BAY' has its own share of the number of records. Below in Figure 10 is a percentage histogram showing the percentage that each of the different levels have for the variable. The different levels from least frequent to highest are: ISLAND, NEAR BAY, NEAR OCEAN, INLAND, and <1H OCEAN. They seem to indicate as the description says, their proximity to the ocean/sea. From lowest to highest the proportions are roughly: 0.0002, 0.1111, 0.1286, 0.3179, and 0.4421. Also, since this is a qualitative variable it's not possible to use ordinary descriptive statistics such as mean, median, or mode.

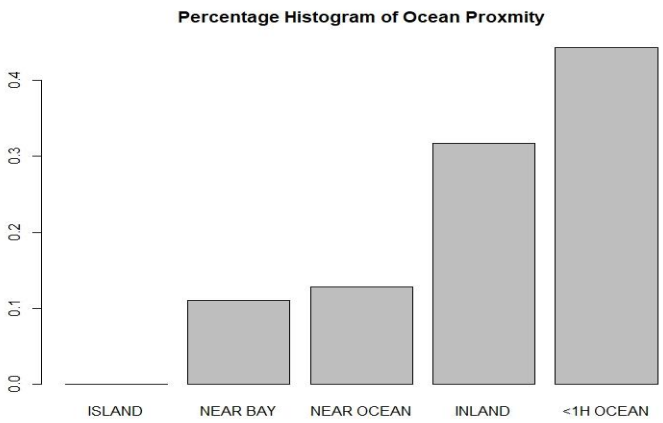

Figure 10

11. The eleventh variable is `class`, this variable was created through feature engineering. It's a qualitative ordinal variable since the levels of lower, middle, and upper signify some sort of ordering from low to high. This variable is based off some analysis done in the previous project with the same dataset. It was found through the website Investopedia that in 1991, 18% of Americans belonged to the lowest class while 5% belong to the highest class. The same statistic was applied ad-hoc to this dataset to get a rough estimate of how the different records could be split into different groups. The `median_income` level was used as the variable to determine which class a record would belong to. If the record was roughly in the bottom 18% of `median_income`, they would be assigned to the lower class. If the record was roughly in the top 5% of `median_income`, they would be assigned to the upper class. The remaining records were assigned to the middle class. In doing so, the following proportions were found for the lower, middle, and upper classes:

Class	*Lower*	*Middle*	*Upper*
Proportion	0.1796918	0.7705535	0.0497547

Also, below in Figure 11 is a percentage histogram showing the same data. It's apparent that the level of 'middle' makes up the largest portion of this variable, followed by 'lower,' and then by 'upper.'

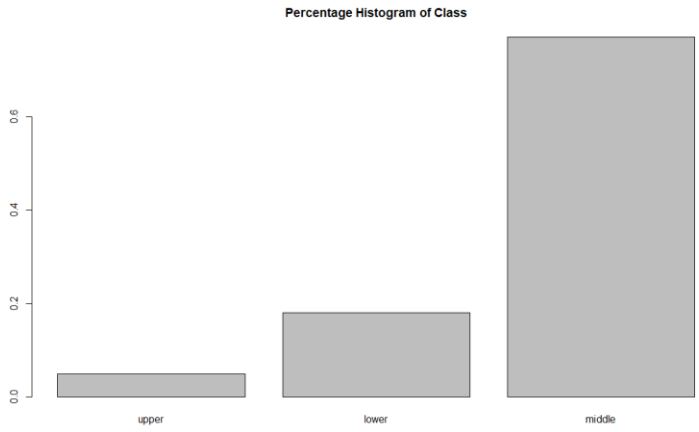

Figure 11

12. The twelfth variable is `city`, this variable was created through feature engineering. It's a qualitative variable of type nominal since order is irrelevant. This variable was created using the `ggmap` package in RStudio. By doing a reverse look up of the city based on a record's latitude and longitude information, a city was assigned a certain value. It's important to note that the accuracy of the latitude and longitude aren't too precise as they only go up to 2 decimal places. Therefore, the accuracy of this variable is not certain. There are 1,050 unique city values and so a specific table showing all of them will not be used. Below in Figure 12 is a barplot of the top 10 cities according to their count within the dataset.

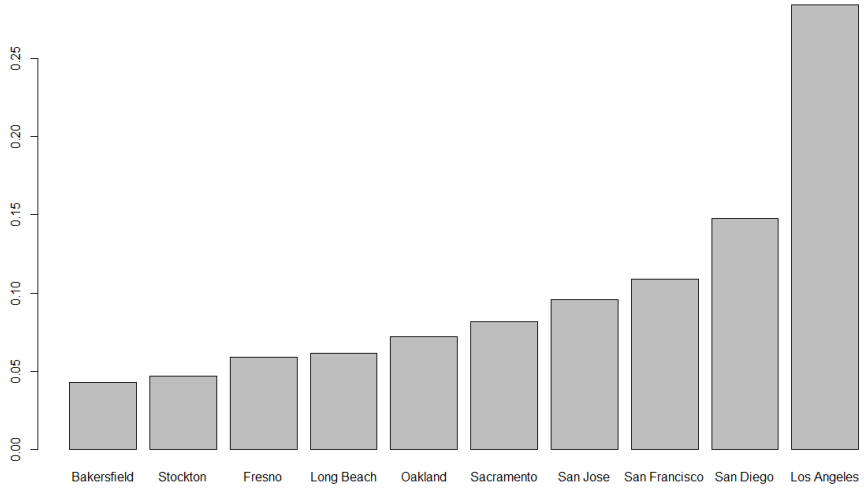

Figure 12

The city of Los Angeles has the highest count with 1,466 records and Bakersfield has the fewest with 221 records. To put it into perspective, there are 15,022 other cities that don't belong in the top 10.

13. The thirteenth variable is `sfla`, this variable was created through feature engineering. It's a qualitative variable of type nominal since order is irrelevant. To create the variable, a list of city names belonging to either the San Francisco Bay Area and the Greater Los Angeles Area were found from the internet. Records were then assigned to 'SF', 'LA', or 'other,' depending which category the fit into. In some situations, there was no city properly assigned to a record and so these records were dropped from the dataset. Also, there are certain records that are clearly within either the San Francisco Bay Area or the Greater Los Angeles are but were assigned a city that didn't belong to either of the group of cities found online. In such situations, they were simply assigned to the level of 'other.' Below is a table showing the proportions for each of these categories. From here, it's interesting to see that almost half the records belong to the Greater Los Angeles Area and around 20% belong to the San Francisco Bay Area. Previously, in project 2, it was evident that the greatest density of records belonged to these areas and so it made sense to focus on them rather than look at the entire state which was sparse due to its large size.

sfla	LA	SF	other
Proportion	0.4395659	0.2025373	0.3578968

Five Analytical Questions

1. What area are most records concentrated in?
2. How to understand the 958 records that all have median house value of $500,001?
3. Where are the most expensive / cheap homes located?
4. Where are the highest / lowest income families?
5. How does the ocean proximity affect the characteristics of a home?

Design

In the previous project it became apparent that to find out information one of the most useful techniques was to implement geospatial graphing of the data. Therefore, it was certain from the beginning that one of the techniques to use would be the geospatial plotting feature in Tableau. A possibility would be to use a single map, however, due to the fact that the majority of records are found in two separate areas of California (i.e., the San Francisco Bay Area and the Greater Los Angeles Area), an idea was to include two separate maps on the dashboard. In one map, the graph would use the search button to focus on San Francisco, California and in the other map the graph would focus on Los Angeles, California. In both maps, some zooming-out would be done to encompass the entirety of both the San Francisco Bay Area and the Greater Los Angeles Area. This can be seen below in Figures 13 and 14. In Figure 13 is the first dashboard where the records represent the `median_income` and in Figure 14 they represent `median_house_value` (next it will be explained why there are two dashboards). The geospatial maps are density maps of the records and don't directly indicate which records have the largest values, but instead focus on the density of the records on the map surface. If a user is interested in seeing the highest/lowest values, these can be subset using global filters which are explained later.

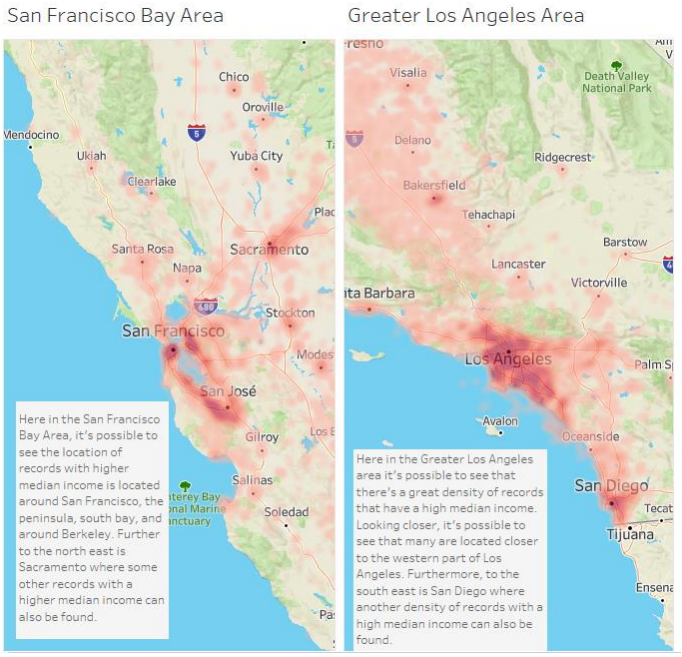

Figure 13

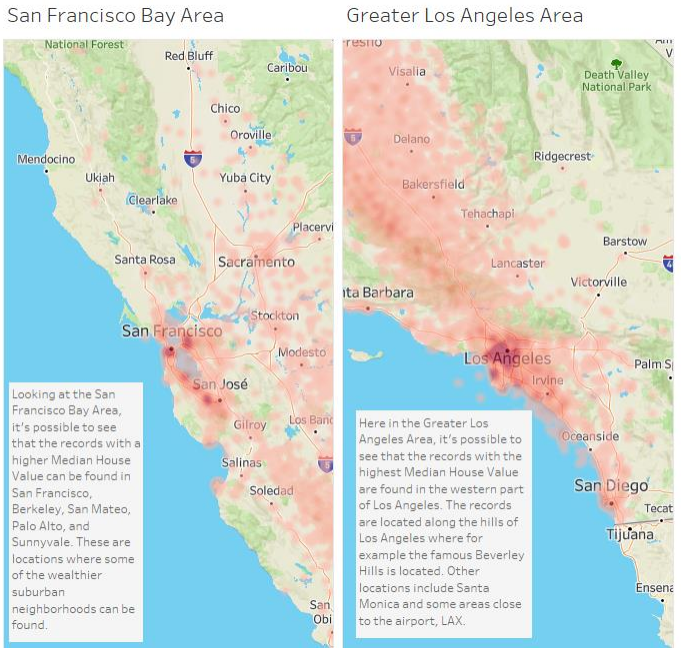

Figure 14

Based on the dataset and the approach towards generating data visualizations, two immediate variables were the focus of interest: `median_income` and `median_house_value`. Therefore, instead of a single dashboard, two of similar template were used where the variable of interest would change between the two. Next, utilizing the story function of Tableau, the two dashboards could be displayed in a story-format where users could analyze one then the other. This can be seen below in Figure 15.

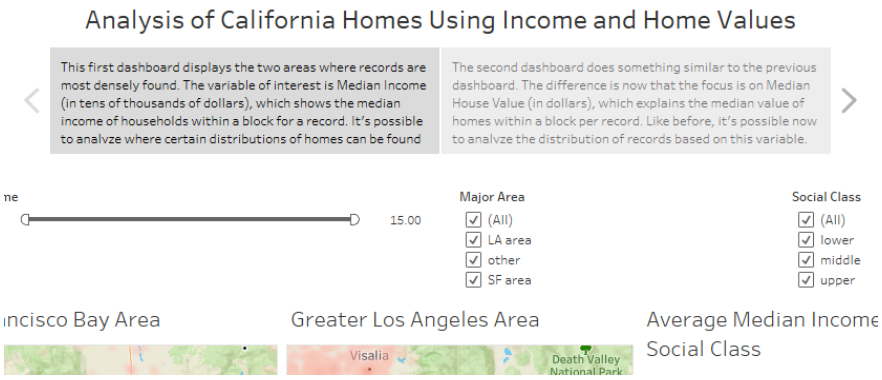

Figure 15

A requirement for this project was that at least two global filters were required for the dashboard. However, when looking through the dataset it didn't seem like there were useful enough filters that could be utilized as global filters. To overcome this, some feature engineering was done to find out the social class and general location of a record. This is done with the variables: `class` and `sfla`. By filtering the records through these two variables, it's possible to subset records based on whether they're located in one of the major areas (i.e., San Francisco Bay Area or the Greater Los Angeles Area) or within a certain social class (i.e., lower class, middle class, or upper class). In this manner, it's possible to subset records so that there can for example be a focus on the upper-class homes or lower-class homes to better understand their location and characteristics. Also, by narrowing down to either the San Francisco Bay Area or the Greater Los Angeles Area, it's possible to get a more precise understanding of those different localities.

In addition to the geospatial plots, one other visualization technique was utilized which are bar charts. However, in the dashboards both vertical and horizontal bar charts were used. The vertical bar chart in this case was used to represent the average median income or average median house value against the social class. The horizontal bar chart can be used to visualize the average median income or average median house value versus the major area. The two different types (i.e., average median income or average median house value) can be seen below in Figures 16 and 17. In both cases, it's possible to see that the upper-class has both a significantly higher average median income and median house value than the other two classes. It's also possible to note that the San Francisco Bay Area and Greater Los Angeles Area have a roughly close average median income and average median house value. However, the San Francisco Bay Area has slightly higher average values for both cases. Records that are outside of these major hubs have noticeably lower values in comparison. Although such information may be intuitive, it's important to have statistics to back up such ideas.

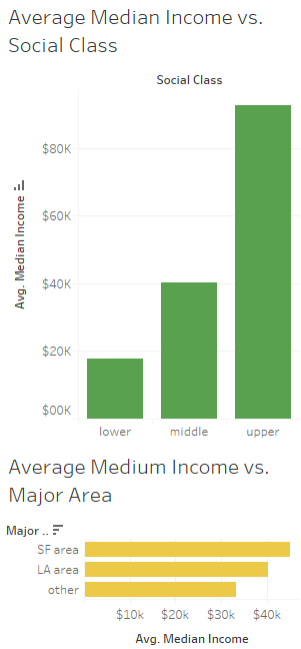

Figure 16

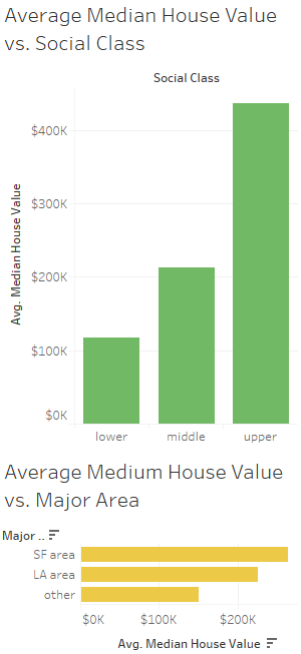

Figure 17

Next are screenshots of both dashboards in Figures 18 and 19. Looking at the two figures it's possible to see how they share a similar template with the only difference being the variable of interest being examined (i.e., either average median income or average median house value). Since there are two dashboards, they were included in a story where there's a brief description to introduce the two dashboards (this can be seen in Figure 15). Furthermore, there are descriptions

that have been placed into each geospatial map to help guide users towards gaining insights that are apparent when analyzing records carefully through the interactive maps. However, people with less knowledge about the areas would likely have a more limited ability to analyze the data that is available. Therefore, the text boxes aid in this manner. In the upper portion are the global filters. The top-left filter is either median income or median house value and the slide bar makes it possible to subset the data towards one of many possible ranges. Since the geospatial graph is that of a density plot, by filtering only the highest or lowest values it's possible to see where the poor or rich are located. The filter in the top-center is used to subset the data by major area. For example, by looking at just the records in the San Francisco Bay Area, it's possible to see their unique social classes matched up against average median income or average median house value. It's apparent from the horizontal bar chart that the two subsets don't have an identical distribution and so having this option allows for a more powerful analysis. The filter on the top-right allows users to subset the data by social class and allows for a similar purpose as when trying to subset by major area. There's potential for a more powerful analysis when trying to analyze specific questions.

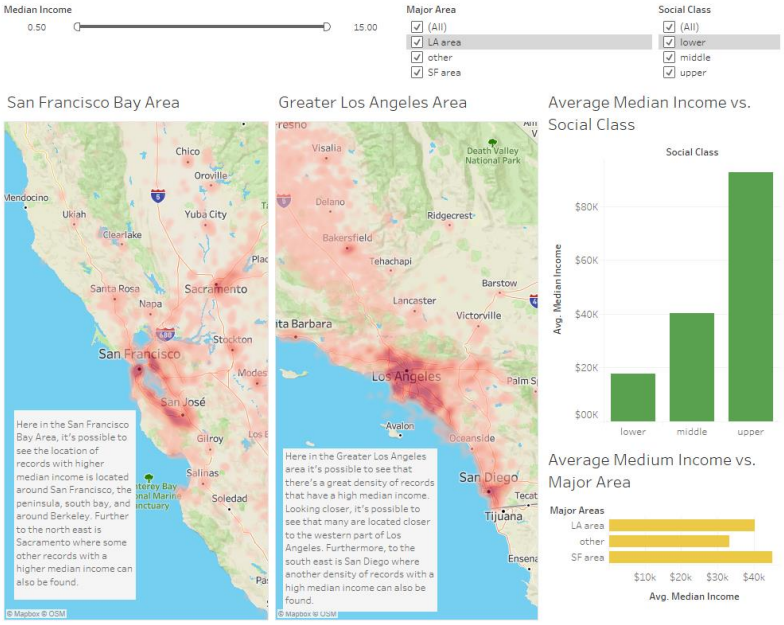

Figure 18

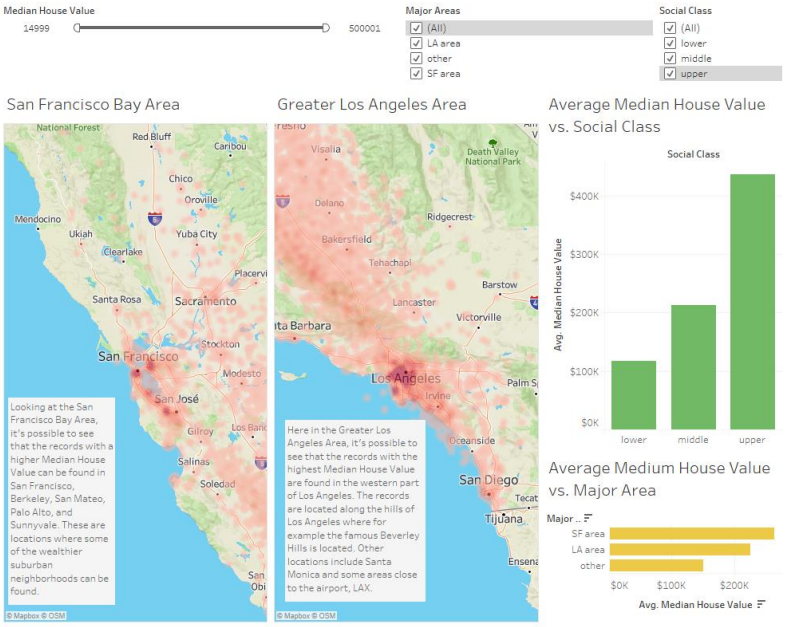

Figure 19

Discussion

The first analytical question asks where most of the records are located. Although the story/dashboards don't directly answer this question, it can be inferred through utilization of the map tools that most of the records are in either the San Francisco Bay Area or the Greater Los Angeles Area. If zooming out of any of the maps is done, it's possible to see the other areas of California where different records are located. For example, in Sacramento of northern California is a small density of records. Also, around Fresno in central California is a sparse density of records too. It becomes apparent that many homes are close to the ocean with the greatest density in the San Francisco Bay Area and the Greater Los Angeles Area along with San Diego. This can be seen below in Figure 20.

Figure 20

Another possibility is to find out where both the most expensive and least expensive homes are located. By using the second dashboard, the variable of interest becomes the median house value of each record. Using the global filter slide bar, it's possible to select any set of range of median home values. For example, looking at homes from $450k - $500k, it's possible to see that the number of records becomes increasingly sparse. Around the San Francisco Bay Area, many are located along the peninsula and some north around San Rafael. In the Greater Los Angeles area, the records become much more focused around the western part of Los Angeles and north around Burbank. This can be seen below in Figure 21.

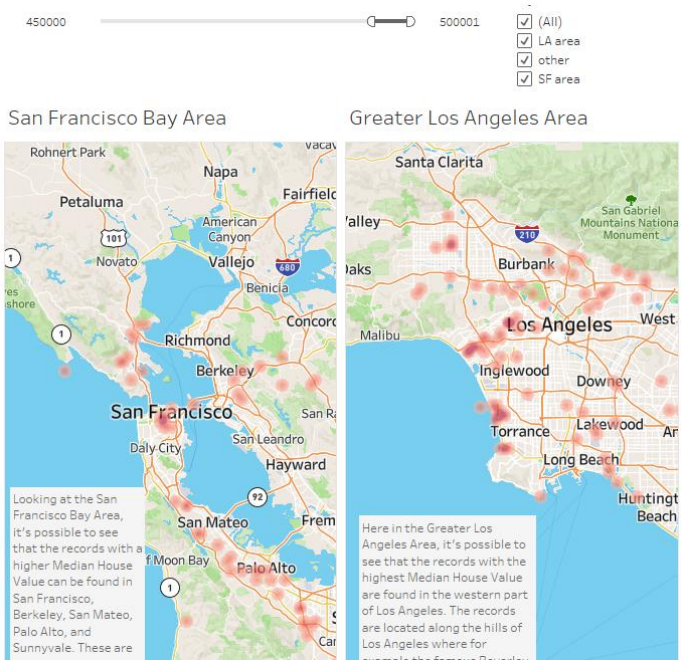

Figure 21

The least expensive homes on the other hand can be seen below in Figure 22. In the following screenshot, the median home values have been subset to homes from around $15k - $200k. In this scenario, the homes around the San Francisco Bay Area that fit within this range are mostly all on the north eastern side of the Bay Area. These include areas such as: Oakland, Berkeley, and Richmond. Within the Greater Los Angeles Area, many of these records can be found around, but there's a much greater density in the center of Los Angeles. This shows that in both cases there's a great split between where the expensive homes and cheaper homes are located. In the latter case, they group in one area and in the former they group in another area.

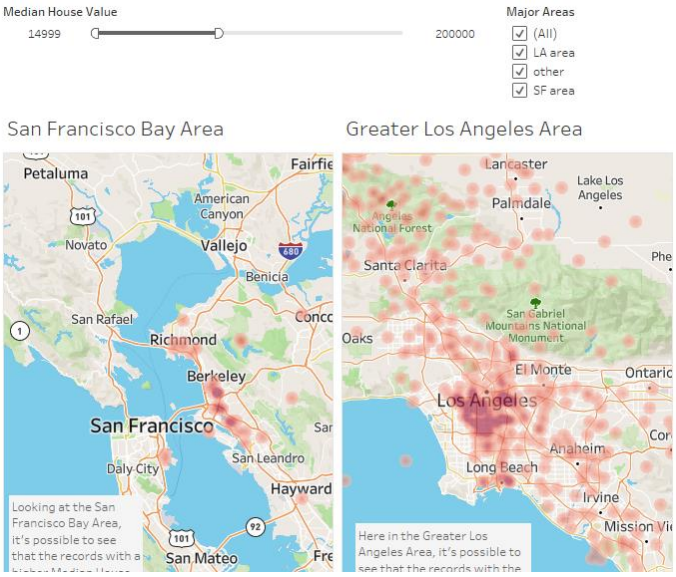

Figure 22

When thinking about where the highest vs. lowest income families are located, it's possible to subset the records by upper and lower social classes. Below in Figure 23 are the homes that belong to the upper class. In the San Francisco Bay Area it's interesting to see that although there is a sparser set of records, they're still scattered throughout the Bay Area. However, in the Greater Los Angeles Area, the upper class seems to distinctly live in areas outside of central Los Angeles. Primarily, they remain in the western and northern areas along with in places such as Irvine.

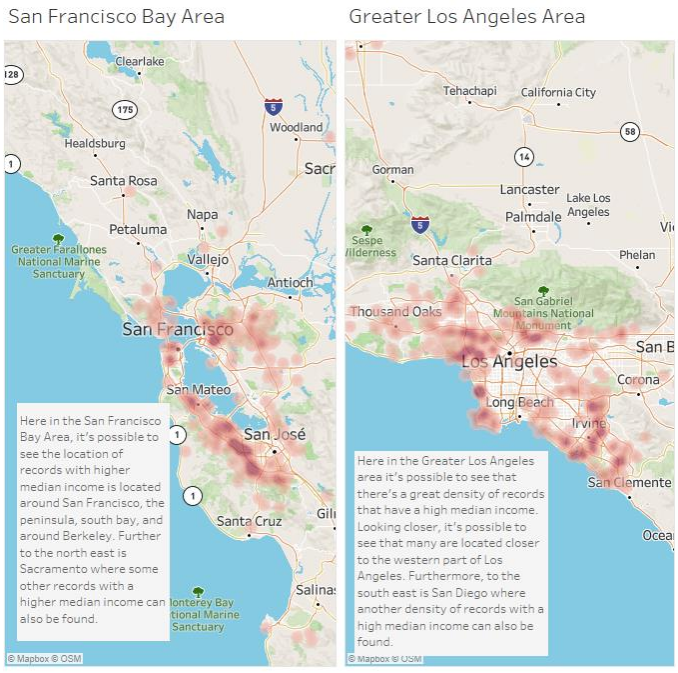

Figure 23

Below in Figure 24 however is the subset of records that belong to the lower class. Around the San Francisco Bay Area, it's possible to see a concentration of them around the eastern part of the city San Francisco along with along the north east and eastern parts of the Bay Area around Richmond, Berkeley, and Oakland.

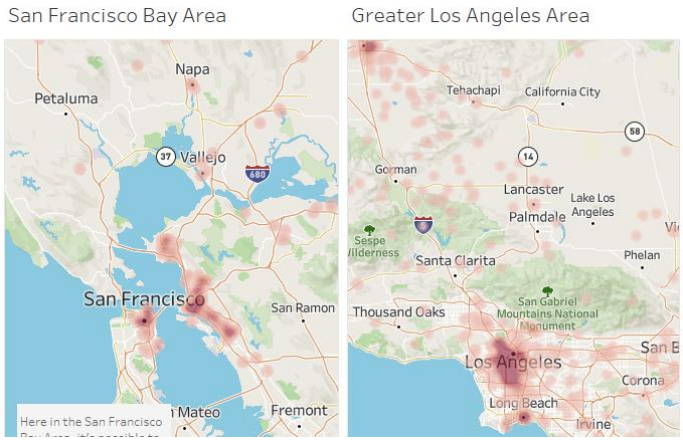

Figure 24

Another analysis question which can be approached is to find the records that have the median home value of $500,001. Interestingly, this problem was solvable quite easily in the previous project. Those records were simply subset in RStudio and plotted using geospatial graphing commands. However, there is some confusion as to why it's not so simple in Tableau. Looking through the records in Tableau, there seem to only be a handful of such records. However, when looking at the same data in RStudio there are over 900 such records. This seems to show some possible limitations that are in place for Tableau with student licenses, or some other reason that is not yet fully understood. Below in Figure 25 are the handful of records with the median home value of $500,001 when using Tableau.

Figure 25

References:

https://www.kaggle.com/camnugent/california-housing-prices

https://www.quora.com/What-was-it-like-to-live-in-Palo-Alto-pre-1990

https://car.sharefile.com/share/view/s0c02663a5c54e23a

https://www.investopedia.com/financial-edge/0912/which-income-class-are-you.aspx

Appendix:

```
library(ggplot2); library(ggmap); library(gridExtra); library(writexl); library(ks)
# Source: https://www.kaggle.com/camnugent/california-housing-prices#housing.csv
housing <- read.csv(file = 'housing.csv')

# check NA's
sum(is.na(housing)) # 207
# Reference: https://stackoverflow.com/questions/4862178/remove-rows-with-all-or-some-nas-
missing-values-in-data-frame
housing <- housing[complete.cases(housing),]
rownames(housing) <- NULL

# Reference: http://blog.chapagain.com.np/r-calculate-mean-median-mode-variance-standard-
deviation/
# Reference: https://community.rstudio.com/t/export-rstudio-data-to-excel/7579/4
descriptive_statistics <- function(x, file_name) {
  iqr_data <- summary(x)
  std_dev <- sd(x)
  unique_x <- unique(x)
  mode <- unique_x[which.max(tabulate(match(x, unique_x)))]
  desc_stat <- as.data.frame(matrix(c(iqr_data, std_dev, mode), ncol = 8))
  colnames(desc_stat) <- c("Minimum", "1st Quantile", "Median", "Mean",
                           "3rd Quantile", "Max", "Std. Deviation", "Mode")
  write_xlsx(x = desc_stat, path = paste(file_name, '.xlsx'), col_names = TRUE)
}

# colnames(housing)
# longitude - A measure of how far west a house is; a higher value is farther west
# quantitative, interval
longitude <- housing[,1]
par(mfrow = c(1,2))
hist(longitude, main = 'Histogram of Longitude', xlab = 'Longitude')
```

```r
boxplot(longitude, main = 'Boxplot of Longitude', ylab = 'Longitude')
median(longitude)
descriptive_statistics(x = longitude, file_name = 'longitude')

# latitude - A measure of how far north a house is; a higher value is farther north
# quantitative, interval
latitude <- housing[,2]
hist(latitude, main = 'Histogram of Latitude', xlab = 'Latitude')
boxplot(latitude, main = 'Boxplot of Latitude', ylab = 'Latitude')
median(latitude)
descriptive_statistics(x = latitude, file_name = 'latitude')

# housing_median_age - Median age of a house within a block; a lower number is a newer
building
# quantitative, ratio
housing_median_age <- housing[,3]
hist(housing_median_age, main = 'Histogram of Median Age of Homes', xlab = 'Median Age')
boxplot(housing_median_age, main = 'Boxplot of Median Age of Homes', ylab = 'Median Age')
min(housing_median_age)
max(housing_median_age)
descriptive_statistics(x = housing_median_age, file_name = 'housing_median_age')

# total_rooms - Total number of rooms within a block
# quantitative, ratio
total_rooms <- housing[,4]
hist(total_rooms, main = 'Histogram of Total Rooms', xlab = 'Total Rooms')
boxplot(total_rooms, main = 'Boxplot of Total Rooms', ylab = 'Total Rooms')
min(total_rooms); max(total_rooms); median(total_rooms)
descriptive_statistics(x = total_rooms, file_name = 'total_rooms')

# total_bedrooms - Total number of bedrooms within a block
# quantitative, ratio
total_bedrooms <- housing[,5]
hist(total_bedrooms, main = 'Histogram of Total Bedrooms', xlab = 'Total Bedrooms')
boxplot(total_bedrooms, main = 'Boxplot of Total Bedrooms', ylab = 'Total Bedrooms')
min(total_bedrooms); max(total_bedrooms); median(total_bedrooms)
descriptive_statistics(x = total_bedrooms, file_name = 'total_bedrooms')

# population - Total number of people residing within a block
# quantitative, ratio
population <- housing[,6]
hist(population, main = 'Histogram of Total Residents', xlab = 'Total Residents')
boxplot(population, main = 'Boxplot of Total Residents', ylab = 'Total Residents')
min(population); max(population); median(population)
descriptive_statistics(x = population, file_name = 'population')

# households - Total number of households, a group of people
# residing within a home unit, for a block
# quantitative, ratio
households <- housing[,7]
hist(households, main = 'Histogram of Total Households', xlab = 'Total Households')
boxplot(households, main = 'Boxplot of Total Households', ylab = 'Total Households')
min(households); max(households); median(households)
descriptive_statistics(x = households, file_name = 'households')

# median_income - Median income for households within a block
```

```r
# of houses (measured in tens of thousands of US Dollars)
# quantitative, ratio
median_income <- housing[,8]
hist(median_income, main = 'Histogram of Median Income', xlab = 'Median Income')
boxplot(median_income, main = 'Boxplot of Median Income', ylab = 'Median Income')
min(median_income); max(median_income); median(median_income)
descriptive_statistics(x = median_income, file_name = 'median_income')

# median_house_value - Median house value for households within
# a block (measured in US Dollars)
# quantitative, ratio
median_house_value <- housing[,9]
hist(median_house_value, main = 'Histogram of Median House Value', xlab = 'Median House
Value')
boxplot(median_house_value, main = 'Boxplot of Median House Value', ylab = 'Median House
Value')
sum(median_house_value == max(median_house_value))
descriptive_statistics(x = median_house_value, file_name = 'median_house_value')

# ocean_proximity - Location of the house w.r.t ocean/sea
# qualitative, nominal
ocean_proximity <- housing[,10]
# Reference: https://stackoverflow.com/questions/21639392/make-frequency-histogram-for-factor-
variables
# barplot(prop.table(table(ocean_proximity)), main = 'Percentage Histogram of Ocean Proxmity')
barplot(prop.table(table(ocean_proximity))[order(prop.table(table(ocean_proximity)))], main =
'Percentage Histogram of Ocean Proxmity')
descriptive_statistics(x = prop.table(table(ocean_proximity)), file_name = 'ocean_proximity')


library(ggmap); library(stringr)

housing <- read.csv(file = 'housing.csv') # Read data

# check NA's
sum(is.na(housing)) # 207
# Reference: https://stackoverflow.com/questions/4862178/remove-rows-with-all-or-some-nas-miss
ing-values-in-data-frame
housing <- housing[complete.cases(housing),]
rownames(housing) <- NULL

# Create class variable
low_to_high_inc <- housing[order(housing$median_income),]
bottom_18 <- round(nrow(housing) * 0.18)
top_5 <- round(nrow(housing) * 0.05)
lower_class <- low_to_high_inc[1:bottom_18,]
lower_class$class <- 'lower'
upper_class <- low_to_high_inc[(nrow(housing) - 1 - top_5):nrow(housing),]
upper_class$class <- 'upper'
middle_class <- low_to_high_inc[(bottom_18 + 1):(nrow(housing) - 2 - top_5),]
middle_class$class <- 'middle'
housing <- rbind(lower_class, middle_class, upper_class)
rownames(housing) <- NULL
barplot(prop.table(table(housing$class))[order(prop.table(table(housing$class)))], main = 'Per
centage Histogram of Class')

# Reference: https://stackoverflow.com/questions/29921605/r-how-to-convert-latitude-and-longit
ude-coordinates-into-an-address-human-rea
revgeocode(c(housing[2,'longitude'], housing[2,'latitude']), output = 'all')$
```

```
address <- mapply(FUN = function(lon, lat)
  revgeocode(c(lon, lat)), housing$longitude, housing$latitude)

housing1 <- housing[1:1000,]
housing2 <- housing[1001:2000,]
housing3 <- housing[2001:3000,]
housing4 <- housing[3001:4000,]
housing5 <- housing[4001:5000,]
housing6 <- housing[5001:6000,]
housing7 <- housing[6001:7000,]
housing8 <- housing[7001:8000,]
housing9 <- housing[8001:9000,]
housing10 <- housing[9001:10000,]
housing11 <- housing[10001:11000,]
housing12 <- housing[11001:12000,]
housing13 <- housing[12001:13000,]
housing14 <- housing[13001:14000,]
housing15 <- housing[14001:15000,]
housing16 <- housing[15001:16000,]
housing17 <- housing[16001:17000,]
housing18 <- housing[17001:18000,]
housing19 <- housing[18001:19000,]
housing20 <- housing[19001:20000,]
housing21 <- housing[20001:20433,]

address1 <- mapply(FUN = function(lon, lat)
  revgeocode(c(lon, lat)), housing1$longitude, housing1$latitude)
address2 <- mapply(FUN = function(lon, lat)
  revgeocode(c(lon, lat)), housing2$longitude, housing2$latitude)
address3 <- mapply(FUN = function(lon, lat)
  revgeocode(c(lon, lat)), housing3$longitude, housing3$latitude)
address4 <- mapply(FUN = function(lon, lat)
  revgeocode(c(lon, lat)), housing4$longitude, housing4$latitude)
address5 <- mapply(FUN = function(lon, lat)
  revgeocode(c(lon, lat)), housing5$longitude, housing5$latitude)
address6 <- mapply(FUN = function(lon, lat)
  revgeocode(c(lon, lat)), housing6$longitude, housing6$latitude)
address7 <- mapply(FUN = function(lon, lat)
  revgeocode(c(lon, lat)), housing7$longitude, housing7$latitude)
address8 <- mapply(FUN = function(lon, lat)
  revgeocode(c(lon, lat)), housing8$longitude, housing8$latitude)
address9 <- mapply(FUN = function(lon, lat)
  revgeocode(c(lon, lat)), housing9$longitude, housing9$latitude)
address10 <- mapply(FUN = function(lon, lat)
  revgeocode(c(lon, lat)), housing10$longitude, housing10$latitude)
address11 <- mapply(FUN = function(lon, lat)
  revgeocode(c(lon, lat)), housing11$longitude, housing11$latitude)
address12 <- mapply(FUN = function(lon, lat)
  revgeocode(c(lon, lat)), housing12$longitude, housing12$latitude)
address13 <- mapply(FUN = function(lon, lat)
  revgeocode(c(lon, lat)), housing13$longitude, housing13$latitude)
address14 <- mapply(FUN = function(lon, lat)
  revgeocode(c(lon, lat)), housing14$longitude, housing14$latitude)
address15 <- mapply(FUN = function(lon, lat)
  revgeocode(c(lon, lat)), housing15$longitude, housing15$latitude)
address16 <- mapply(FUN = function(lon, lat)
  revgeocode(c(lon, lat)), housing16$longitude, housing16$latitude)
address17 <- mapply(FUN = function(lon, lat)
  revgeocode(c(lon, lat)), housing17$longitude, housing17$latitude)
address18 <- mapply(FUN = function(lon, lat)
  revgeocode(c(lon, lat)), housing18$longitude, housing18$latitude)
```

```r
address19 <- mapply(FUN = function(lon, lat)
  revgeocode(c(lon, lat)), housing19$longitude, housing19$latitude)
address20 <- mapply(FUN = function(lon, lat)
  revgeocode(c(lon, lat)), housing20$longitude, housing20$latitude)
address21 <- mapply(FUN = function(lon, lat)
  revgeocode(c(lon, lat)), housing21$longitude, housing21$latitude)

address <- c(address1, address2, address3, address4, address5,
      address6, address7, address8, address9, address10,
      address11, address12, address13, address14, address15,
      address16, address17, address18, address19, address20,
      address21)

# Reference: https://stackoverflow.com/questions/45289000/split-a-address-string-into-city-sta
te-and-address-in-r
address_split <- str_match(address,"(.+), (.+), (.+) (.+), (.+)")[ ,-1]
cities <- address_split[,2]
housing$city <- cities
top10_city <- sort(table(housing$city), decreasing = TRUE)[1:10]
top10_city <- housing[housing$city %in% names(top10_city),'city']
table(top10_city)
other_city <- housing[!(housing$city %in% unique(top10_city)),'city']
length(other_city)
# city_freq <- c(top10_city, rep('other', length(other_city)))
barplot(sort(prop.table(table(top10_city))))

# Reference: https://en.wikipedia.org/wiki/San_Francisco_Bay_Area
sf_bay_area <- c('Alameda', 'Albany', 'American Canyon', 'Antioch', 'Atherton',
                 'Belmont', 'Belvedere', 'Benicia', 'Berkeley', 'Brentwood',
                 'Brisbane', 'Burlingame', 'Calistoga', 'Campbell', 'Clayton',
                 'Cloverdale', 'Colma', 'Concord', 'Corte Madera', 'Cotati',
                 'Cupertino', 'Daly City', 'Danville', 'Dixon', 'Dublin',
                 'East Palo Alto', 'El Cerrito', 'Emeryville', 'Fairfax', 'Fairfield',
                 'Foster City', 'Fremont', 'Gilroy', 'Half Moon Bay', 'Hayward',
                 'Healdsburg', 'Hercules', 'Hillsborough', 'Lafayette', 'Larkspur',
                 'Livermore', 'Los Altos', 'Los Altos Hills', 'Los Gatos', 'Martinez',
                 'Menlo Park', 'Mill Valley', 'Millbrae', 'Milpitas', 'Monte Sereno',
                 'Moraga', 'Morgan Hill', 'Mountain View', 'Napa', 'Newark',
                 'Novato', 'Oakland', 'Oakley', 'Orinda', 'Pacifica',
                 'Palo Alto', 'Petaluma', 'Piedmont', 'Pinole', 'Pittsburg',
                 'Pleasant Hill', 'Pleasanton', 'Portola Valley', 'Redwood', 'Richmond',
                 'Rio Vista', 'Rohnert Park', 'Ross', 'St. Helena', 'San Anselmo',
                 'San Bruno', 'San Carlos', 'San Francisco', 'San Jose', 'San Leandro',
                 'San Mateo', 'San Pablo', 'San Rafael', 'San Ramon', 'Santa Clara',
                 'Santa Rosa', 'Saratoga', 'Sausalito', 'Sebastopol', 'Sonoma',
                 'South San Francisco', 'Suisun City', 'Sunnyvale', 'Tiburon', 'Union City',
                 'Vacaville', 'Vallejo', 'Walnut Creek', 'Windsor', 'Woodside',
                 'Yountville')

# Reference: https://support.crunchbase.com/hc/en-us/articles/360009895834-What-cities-are-in-
the-Greater-Los-Angeles-region-
greater_la_area <- c('Los Angeles', 'Santa Monica', 'Irvine', 'Beverly Hills', 'Lucerne Valley
',
                     'Pasadena', 'Newport Beach', 'West Hollywood', 'Culver City', 'El Segundo
',
                     'Venice', 'Torrance', 'Costa Mesa', 'Burbank', 'Santa Ana',
                     'Aliso Viejo', 'Woodland Hills', 'Long Beach', 'Anaheim', 'Westlake Villa
ge',
                     'Glendale', 'Manhattan Beach', 'Marina Del Rey', 'Sherman Oaks', 'Encino
',
                     'Huntington Beach', 'Orange', 'Van Nuys', 'Lake Forest', 'San Clemente',
```

```
                    'Chatsworth', 'Riverside', 'Valencia', 'Calabasas', 'Walnut',
                    'Agoura Hills', 'Redondo Beach', 'Tustin', 'Laguna Hills', 'Thousand Oaks
',
                    'Temecula', 'Fountain Valley', 'Mission Viejo', 'Malibu', 'Studio City',
                    'North Hollywood', 'Laguna Beach', 'Ventura', 'Fullerton', 'Hermosa Beach
',
                    'Murrieta', 'Corona', 'Pomona', 'San Juan Capistrano', 'Rancho Santa Marg
arita',
                    'Rancho Cucamonga', 'Monrovia', 'Gardena', 'Simi Valley', 'Camarillo',
                    'Oxnard', 'Brea', 'Garden Grove', 'Palm Springs', 'Whittier',
                    'Monterey Park', 'Cerritos', 'Palm Desert', 'City Of Industry', 'San Bern
ardino',
                    'Tarzana', 'Buena Park', 'Northridge', 'Pacific Palisades', 'La Canada Fl
intridge',
                    'Canoga Park', 'Ontario', 'Diamond Bar', 'West Covina', 'Laguna Niguel',
                    'Foothill Ranch', 'Playa Vista', 'La Mirada', 'Santa Clarita', 'Claremont
',
                    'Alhambra', 'Lancaster', 'Yorba Linda', 'Redlands', 'Westminster',
                    'Toluca Lake', 'Hawthorne', 'Carson', 'Ladera Ranch', 'Newbury Park',
                    'Rancho Palos Verdes', 'Sylmar', 'South Pasadena', 'Inglewood', 'Arcadia
',
                    'Azusa', 'Glendora', 'Moreno Valley', 'Moorpark', 'Cypress',
                    'Dana Point', 'El Monte', 'Reseda', 'Altadena', 'Palmdale',
                    'Santa Fe Springs', 'Chino', 'San Dimas', 'Panorama City', 'Rancho Doming
uez',
                    'Commerce', 'Fontana', 'Seal Beach', 'Granada Hills', 'Sun Valley',
                    'Playa Del Rey', 'Canyon Country', 'Rosemead', 'Universal City', 'Century
 City',
                    'Covina', 'Upland', 'Chino Hills', 'Victorville', 'Signal Hill',
                    'Tujunga', 'West Hills', 'Artesia', 'Placentia', 'Palos Verdes Estates',
                    'Lakewood', 'San Fernando', 'Norwalk', 'Lomita', 'South El Monte',
                    'Compton', 'Temple City', 'Baldwin Park', 'West Los Angeles', 'Valley Vil
lage',
                    'San Pedro', 'Bellflower', 'La Puente', 'La Crescenta', 'Huntington Park
',
                    'Montebello', 'Rowland Heights', 'Colton', 'Wildomar', 'Lake Elsinore',
                    'Yucca Valley', 'La Habra', 'Sunland', 'Mission Hills', 'South Gate',
                    'Downey', 'Rolling Hills Estates', 'Wilmington', 'San Marino', 'San Gabri
el',
                    'Lynwood', 'Barstow', 'Yucaipa', 'Norco', 'Newport Coast',
                    'Cathedral City', 'La Quinta', 'Desert Hot Springs', 'Rancho Mirage', 'Oj
ai',
                    'Pacoima', 'North Hills', 'Oak Park', 'Stanton', 'La Palma',
                    'Maywood', 'Duarte', 'Paramount', 'Pico Rivera', 'Stevenson Ranch',
                    'Menifee', 'Hesperia', 'Montclair', 'Loma Linda', 'Apple Valley',
                    'Grand Terrace', 'Beaumont', 'Trabuco Canyon', 'Laguna Woods', 'Indio',
                    'Hemet', 'San Jacinto', 'Santa Paula', 'Acton', 'Harbor City',
                    'Joshua Tree', 'Mountain Pass', 'Villa Park', 'Blythe', 'Adelanto',
                    'Bell Gardens', 'Newhall', 'Montrose', 'Palos Verdes Peninsula', 'Topanga
',
                    'Lawndale', 'Highland', 'Mira Loma', 'Lake Arrowhead', 'La Verne',
                    'Sun City', 'Hacienda Heights', 'Rialto', 'Los Alamitos', 'Idyllwild',
                    'Indian Wells', 'Coachella', 'Banning', 'Midway City', 'Sierra Madre',
                    'Thermal', 'Port Hueneme Cbc Base', 'Sunset Beach', 'Silverado', 'Avalon
',
                    'Castaic', 'East Los Angeles', 'Bell', 'Bloomington', 'Bryn Mawr',
                    'Helendale', 'Baker', 'Mentone', 'Big Bear City', 'Patton',
                    'Guasti', 'Big Bear Lake', 'Perris', 'Calimesa', 'Hawaiian Gardens',
                    'Capistrano Beach', 'Twentynine Palms', 'Thousand Palms', 'Cabazon', 'Fil
lmore',
                    'Port Hueneme', 'Ludlow', 'Running Springs', 'Lytle Creek', 'North Long B
```

```
each',
                    'Oro Grande', 'Nuevo', 'Landers', 'Phelan', 'Parker Dam',
                    'Winnetka', 'Rimforest', 'Newberry Springs', 'Verdugo City', 'Pioneertown
',
                    'Crestline', 'Blue Jay', 'March Air Reserve Base', 'Valyermo', 'Morongo V
alley',
                    'Trona', 'Charter Oak', 'Somis', 'Mount Wilson', 'Point Mugu Nawc',
                    'Forest Falls', 'Littlerock', 'Anza', 'Mecca', 'Winchester',
                    'Amboy', 'Crest Park', 'Wrightwood', 'Piru', 'Atwood',
                    'Cedar Glen', 'El Toro', 'Fort Irwin', 'Angelus Oaks', 'Skyforest',
                    'Vidal', 'Cedarpines Park', 'Rolling Hills', 'Oak View', 'Mountain Center
',
                    'Earp', 'Hinkley', 'Surfside', 'Fawnskin', 'Yermo',
                    'Aguanga', 'North Palm Springs', 'Daggett', 'Green Valley Lake', 'Pearblo
ssom',
                    'Twin Peaks', 'Sugarloaf', 'Llano', 'Nipton', 'East Irvine',
                    'Desert Center', 'Lake Hughes', 'Brandeis', 'Homeland', 'Needles')

# Reference: https://community.rstudio.com/t/adding-column-based-on-other-column/16114/2
housing$sfla <- ifelse(housing$city %in% sf_bay_area, 'SF',
                       ifelse(housing$city %in% greater_la_area, 'LA','other'))

housing <- housing[complete.cases(housing),]
table(housing$sfla) / sum(table(housing$sfla))

write.csv(x = housing, file = 'C:\\Users\\qizhe\\Desktop\\JHU\\data_vis\\project_3\\housing_up
date.csv', row.names = FALSE)
```