Jared Yu

Data Visualization

Project 2

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

The Data

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.

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.

The Variables

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.

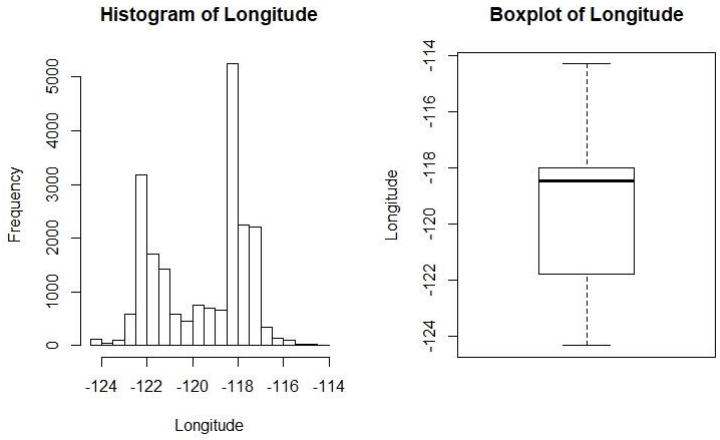

Figure 1

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.

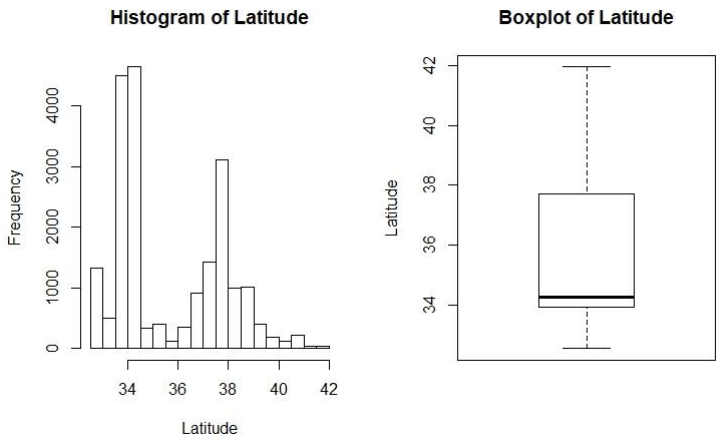

Figure 2

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.

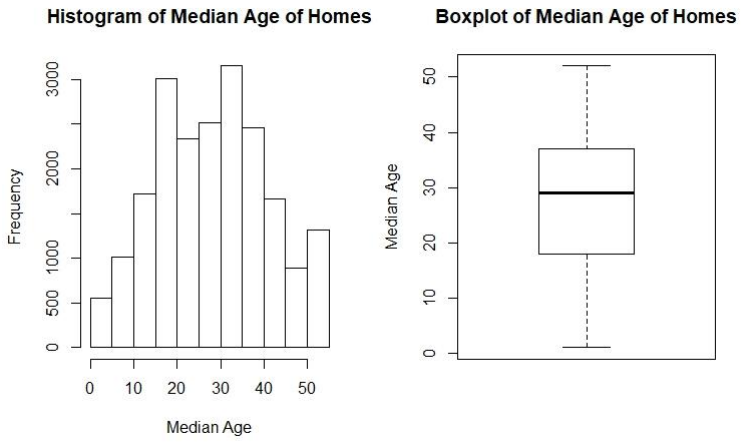

Figure 3

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.

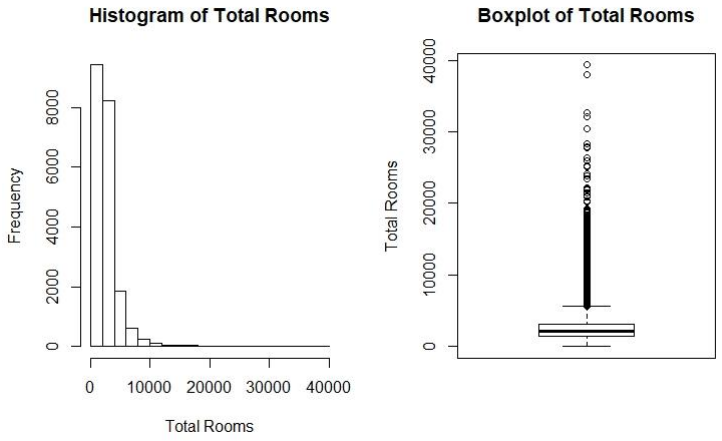

Figure 4

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.

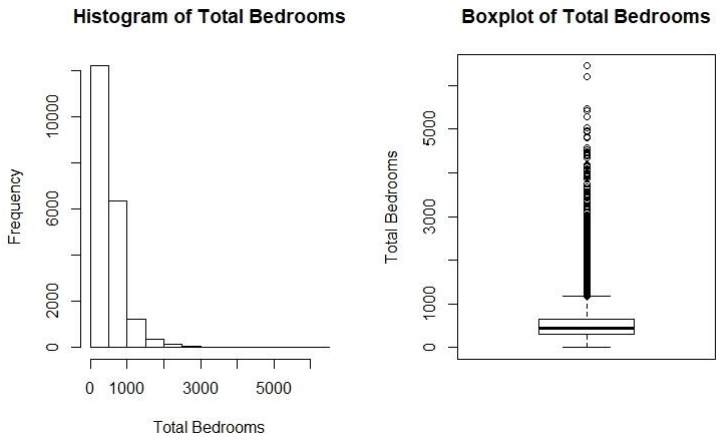

Figure 5

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.

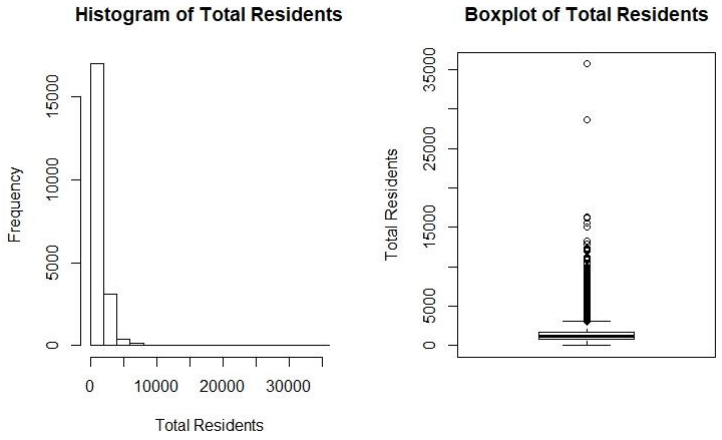

Figure 6

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.

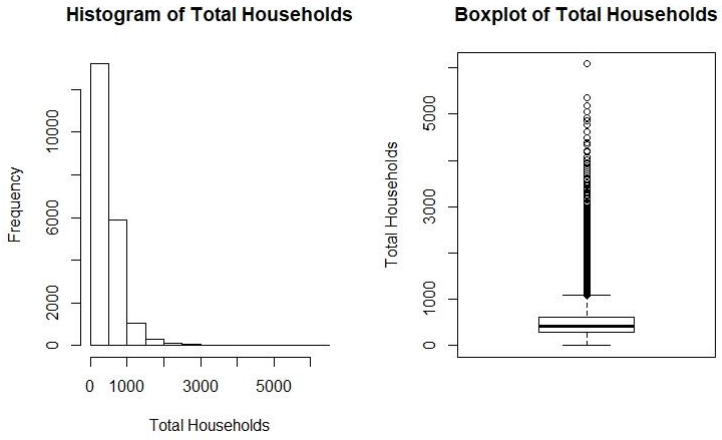

Figure 7

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.

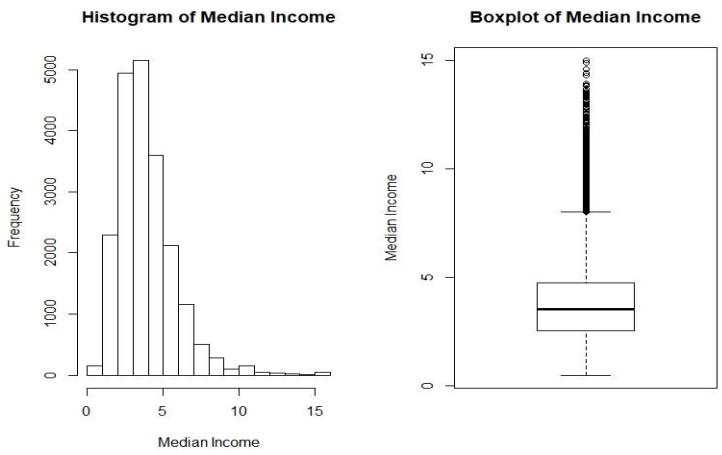

Figure 8

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.

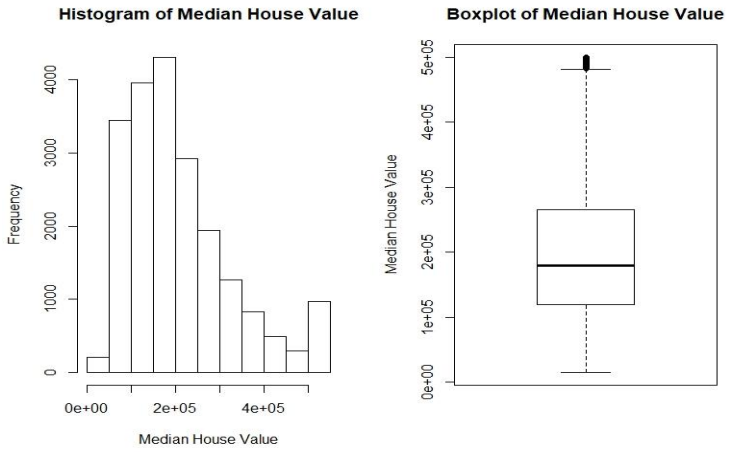

Figure 9

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.

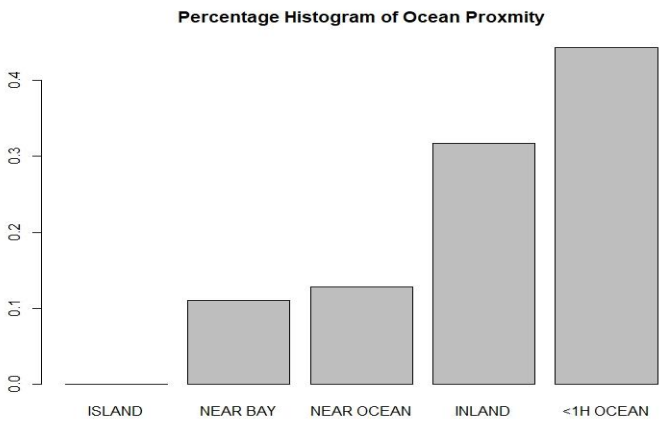

Figure 10

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?

Analysis of Questions

1. The first question to be analyzed with visualizations is, “What area are most records concentrated in?” The most straightforward method to analyze this question is simply to analyze the longitude and latitude of the records and graph them using a geospatial plot. This would assign a point to each record on a geographic map of California. This can be seen in the first visualization as the left image of Figure 11. Using a geographic map such as this one, it’s possible to see that each of the records are in various regions across the state of California with varying densities. This geographic plot makes it such that it’s possible for someone doing analysis to instantly see each and everyone record plotted by their longitude and latitude so that it’s possible to quickly gain understanding towards how the records are spread out geographically.

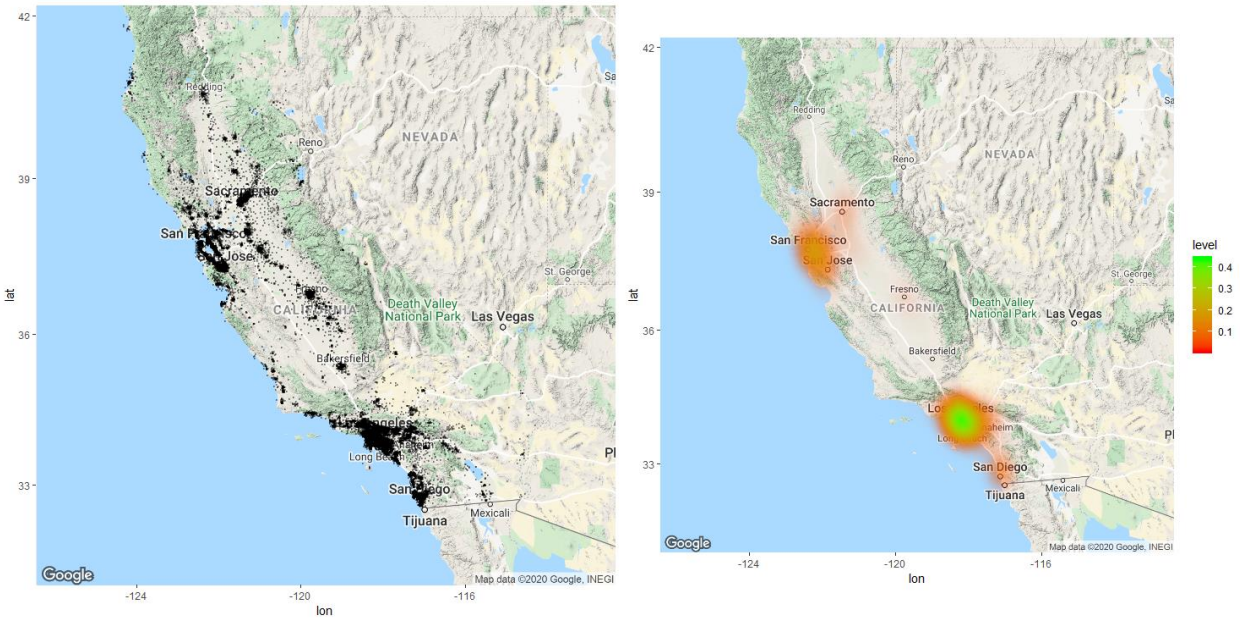

Figure 11

The second visualization shown as the graph on the right-hand side of Figure 11 is a geospatial heatmap. It builds on top of the previous visualization by changing the records denoted as points into a heatmap. The heatmap has the purpose of better explaining the density of the records in a geospatial context. When the records are plotted as points, it's possible to see all of them at once, but it's difficult to understand areas with a high density as they all became a mass of color without distinction from one another. Therefore, by using the heatmap function it's possible to see their density as sort of like a contour map. The highest density is in green, with the lower density in red. It becomes apparent however that less dense regions noticeable previously are no longer seen on the map. By analyzing the heatmap, it's possible to see that there are two major densities in northern and southern California. The densest area however is located around Los Angeles.

The third visualization is seen below in Figure 12. The visualization is a 3D graph of the kernel density estimate (KDE) for the `longitude` and `latitude` variables. The kernel functions and bandwidths for the KDE are default because it's not necessary in this case to become too specific about their hyperparameters. In the following third visualization, KDE helps to better compare the multiple densities that are apparent in the first visualization and to try and see their relative ranking. In the KDE below, it's clear that Los Angeles is the densest region with the major spike. However, it seems that the San Francisco Bay Area and San Diego area have a similar density with the exception that San Diego has a smaller area. Lastly, Sacramento seems to be comparatively the least dense of the major regions previously noted. With this third visualization, KDE is used to

generate a type of 3-Dimensional smoothed histogram that makes it possible to see more clearly and with greater detail what was being plotted in the heatmap.

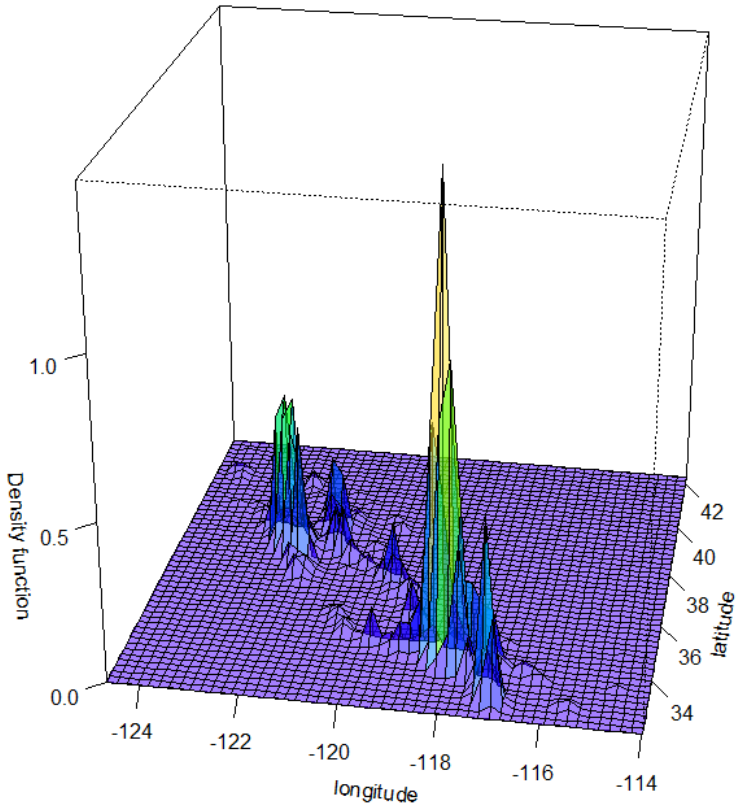

Figure 12

2. The next question to be investigated is, "How to understand the 958 records that all have median house value of $500,001?" The first visualization is based off the most instinctive step which is to subset the 958 records that have a median house value of $500,001 and to plot their coordinates in a geospatial graph like what was seen in Figure 11. The graph of these coordinates can be seen in Figure 13. The thinking is that since there are 958 observation with the same maximum median house value, a basic way to understand them is to see where the homes are located. In urban areas such as the San Francisco Bay Area and the Los Angeles area exists greater economic movement, something that makes it possible for the existence of high-priced housing. In Figure 13, it's evident that the largest concentration is around these areas. The more interesting takeaway is that there are few other homes with the same value outside of these areas, and those that are outside the area aren't too far away.

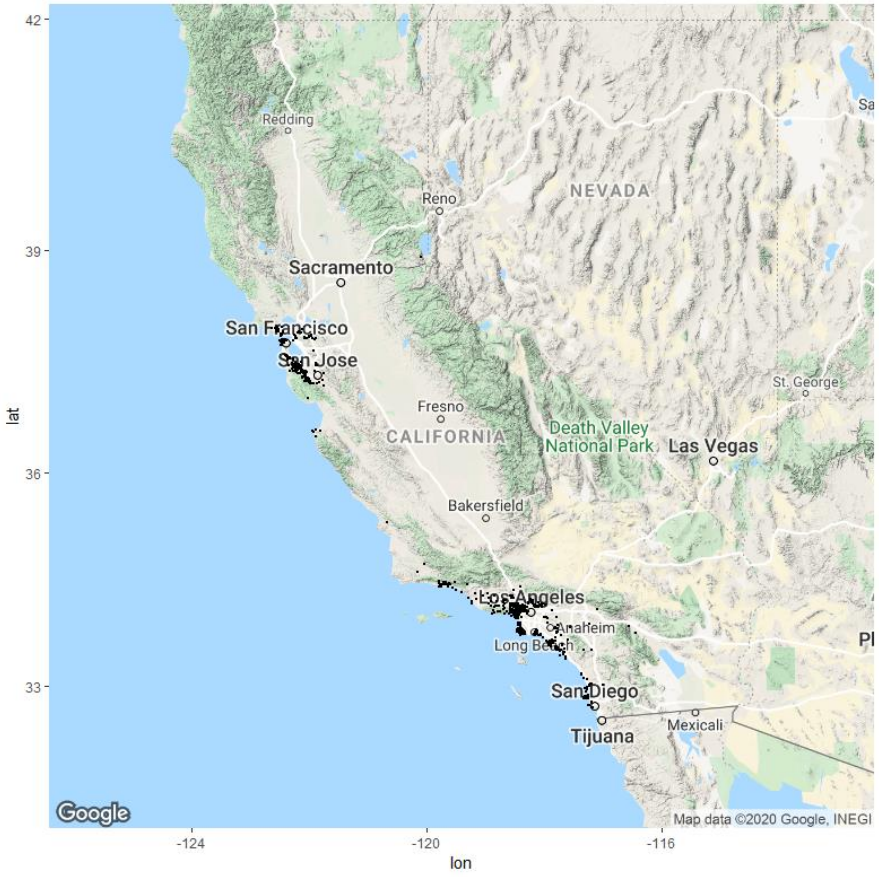

Figure 13

The next step is to zoom in on these locations and try to better see up close the locations where the high-valued homes are located. Below in Figure 14 is the second visualization that shows side-by-side the two areas of San Francisco and Los Angeles. The red dots indicate the high-valued homes and by zooming in it's possible to see more clearly which cities and neighborhoods contain those observations. Looking closer at the plot on the left of Figure 14 makes it clear that the areas: Mill Valley, San Francisco, San Mateo, Palo Alto, and Mountain View contain the largest portion of high-valued homes. Looking at the plot on the right of Los Angeles shows that most points are along the hills of Los Angeles in wealthy neighborhoods such as Beverley Hills. Areas such as this contain mansions that give reason as to why there are such high-valued homes in these areas. After zooming in and finding which neighborhoods contain the high-valued homes, it's possible to do a Google search to find out what those areas were like in the 1990's to confirm the ideas being presented in the visualizations. An interesting note also is that some observations lie in the water, which shows that the coordinates aren't accurate enough at up to two decimal places.

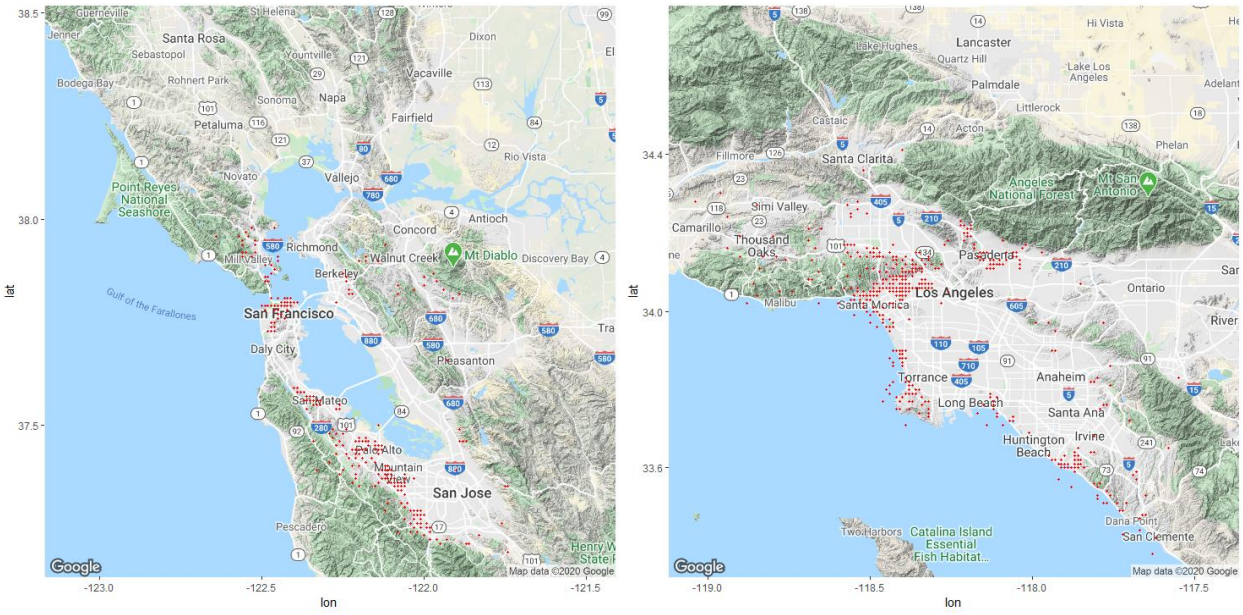

Figure 14

The third and last visualization can be seen in Figure 15 for the second question regarding the 958 homes with a median value of $500,001. It shows clockwise from the top-left are: KDE of `total_rooms`, KDE of `median_income`, percentile histogram of `ocean_proximity`, and KDE of `housing_median_age`. Once again, the kernel functions and bandwidths are set to default as there isn't a great need for tuning those parameters in this visual analysis. The thinking here is that this subset of homes a unique set of characteristics that would differentiate them from the other records in the dataset. By analyzing and comparing them with the overall distributions for each of the variables, it's possible to begin to understand the statistics that make these homes unique. For example, with the KDE of `total_rooms`, it's possible to see that it exhibits a pattern like what's seen in the general dataset but the skew is less extreme. This is a logical result since expensive homes can often be situated in secluded neighborhoods rather than dense urban areas where cheaper homes are often found. When looking at the KDE of `median_income`, it's possible to see that with this subset that the distribution is shifted further to the right. This is an intuitive result since people with expensive homes would need a higher income than average to afford living there. By looking at the percentile histogram of `ocean_proximity`, it's possible to see what was depicted in Figure 14 with most of the homes being located near water. However, the interesting part is that few expensive homes are located inland in comparison to the general dataset where inland is the second most common ocean proximity. Lastly, looking at the KDE of `housing_median_age`, shows that these high-valued homes tend to be older on average than the general population. An interesting reason could be that the value is based largely on the location of the homes. Therefore, homes built in these specific high-valued areas

will tend to appreciate well with time and don't necessarily need to be particularly new or modern.

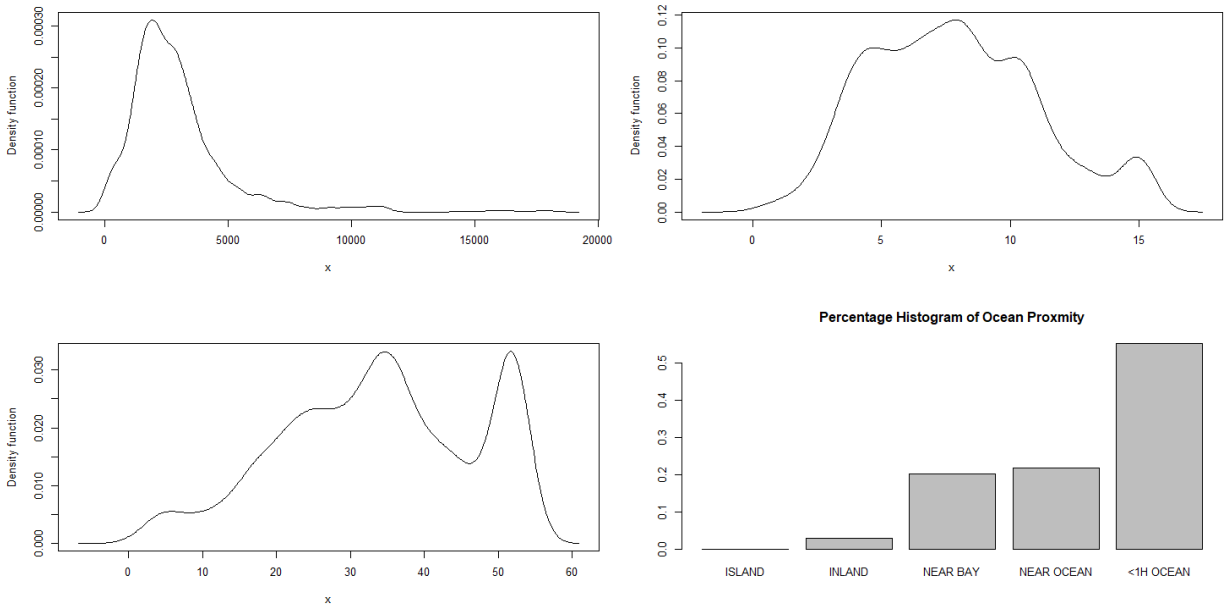

Figure 15

3. The last question to be analyzed is, "Where are the highest / lowest income families?" According to Investopedia in 1991 18% of Americans belonged to the lowest class while 5% belong to the highest class. To understand these two demographics, the dataset was subset according to the above statistic. The lower-class contains the 3,678 observations with lowest median income and the upper-class contains the 1,024 highest median income observations.

   In the first visualization for this question, an instinctive step is to simply plot the coordinates on a map for each of the subsets and compare the differences in their corresponding locations. In Figure 16, the upper-class homes are plotted in blue, while the lower-class homes are plotted in red. On the same geospatial map of California, it's possible then to understand how the distribution of the different subsets pan out across the state. After having gone over the subset of 958 homes that are of extremely high value, here there's a similar pattern of upper-class homes concentrated in those same areas. In the map, it's possible to see how the San Francisco Bay Area and Los Angeles area contain most blue dots, with only a few other select locations having a scarce number of blue dots. On the contrary, red dots are seen intermingled in areas where blue dots exist in addition to being spread out further inland in denser areas such as Sacramento, Fresno, and Bakersfield. Also, in further inland and some places near the ocean are sparse amounts of red dots showing that lower-class homes exist much more evenly throughout the entire state. A logical conclusion is that where there are large economic hubs such as the San Francisco Bay Area or Los Angeles area there will

inevitably be a mixture of rich and poor homes. However, in areas further from economic hubs there is largely a density of poor homes with little wealth being able to concentrate.

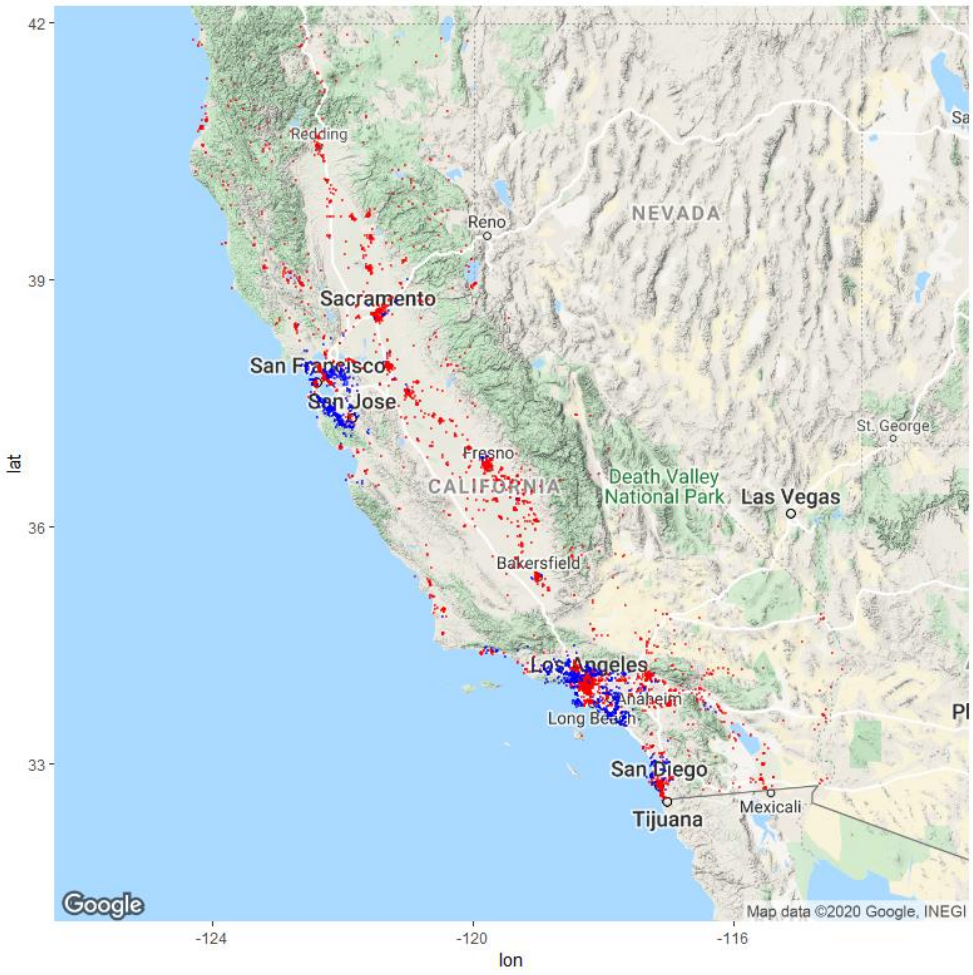

Figure 16

In the second visualization below in Figure 17 is a percentile histogram of the `ocean_proximity` variable for the two subsets of lower-class and upper-class homes. The lower-class homes can be seen on the left-hand side and the upper-class homes on the right-hand side. The reason that this variable is being analyzed for each of the two subsets is that from the previous visualization in Figure 16 it was clear that there was a different distribution of this characteristic for the two subsets. By analyzing their corresponding distributions side-by-side it's possible to see clearly how different they are regarding their proximity to water. With the lower-class homes, in fact the majority are located inland and those near water. Using R, this number is approximately 49.08% or roughly half. In contrast, only approximately 9.28% of upper-class homes are located inland. Such a statistic is quite extreme and shows the intense disparity between lower and upper-class home locations within California. Such information indicates that there is

much weaker economic growth in areas inland in comparison to places next to the ocean or bay. This information could be valuable to state government officials who are interested in finding new ways to generate wealth for the economy. Perhaps it'd be possible for them to design ways to encourage investment in areas further inland to help boost the local economies and diversify away from already highly busy parts of the state.

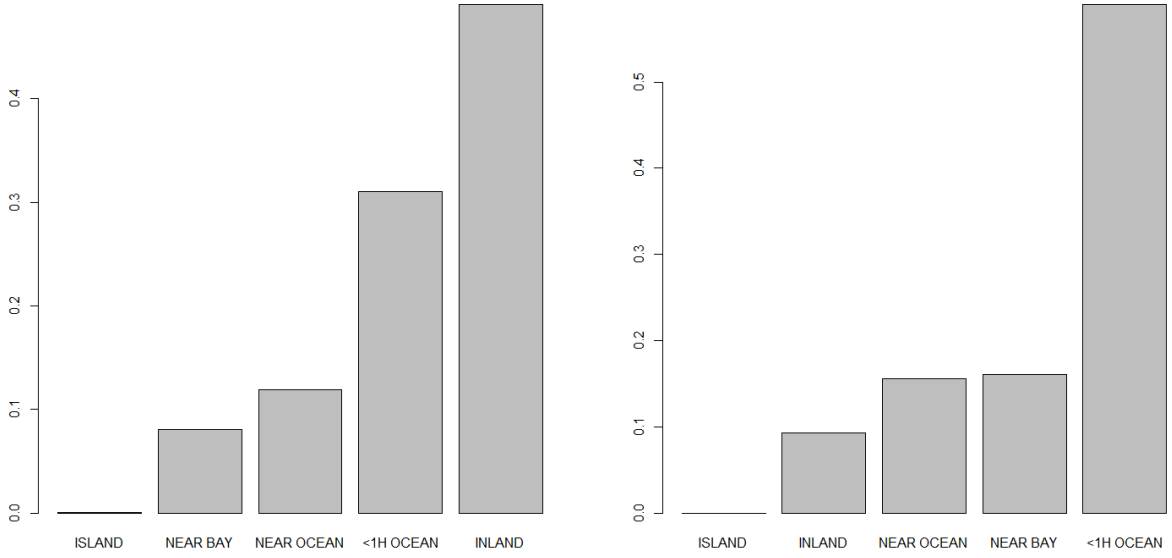

Figure 17

The last visualization for the last question can be seen below in Figure 18. This visualization is a 2×2 grid of histograms and boxplots analyzing the `population` variable for both the lower-class and upper-class. By creating a such a visualization, it's possible to compare both the two together in terms of this variable in addition to comparing them with the overall dataset. The reason why this specific variable is being chosen is because it builds off the previous visualizations that made note of how different the areas could be regarding where these two subsets are located. In the previous visualizations of Figure 16 and 17, it was possible to see how the two subsets are spread in different areas. By analyzing the `population` variable, it's possible to look closely at the density of people within the areas of these homes. In Figure 18, on the upper-half shows the histogram and boxplot for the lower-class and on the lower-half shoes the histogram and boxplot for the upper-class. It's apparent that they both show a similar pattern in both the histogram and boxplot. This is also like what was seen with the general population. However, when looking closer at the numbers it is apparent lower-class homes are more centered around an area of around 1,000, confirmed in R by seeing that their median is 1,128. With upper-class homes, their median is 989, but the skew to the right is much more extreme. Looking at the boxplots, it's possible to see that a few outliers are in fact pulling both distributions much further to the right making it difficult to understand the general overview. It would be possible to remove these outliers and

replot to see another view of the distribution for this variable. Another interesting insight could be that upper-class homes tend to be located within busy economic areas, but they're simultaneously located in more secluded suburban locations with fewer residents per block. This creates a balance in the distribution for upper-class homes. However, lower-class homes have significantly more homes located in less dense inland areas where the overall population is much lower. Yet, in lower-class locations, the homes could be positioned in such a way that within a block there would be more residents. This makes sense from a real estate point of view, since for poor homes they would own much less land and live closer together despite being in overall less dense cities.

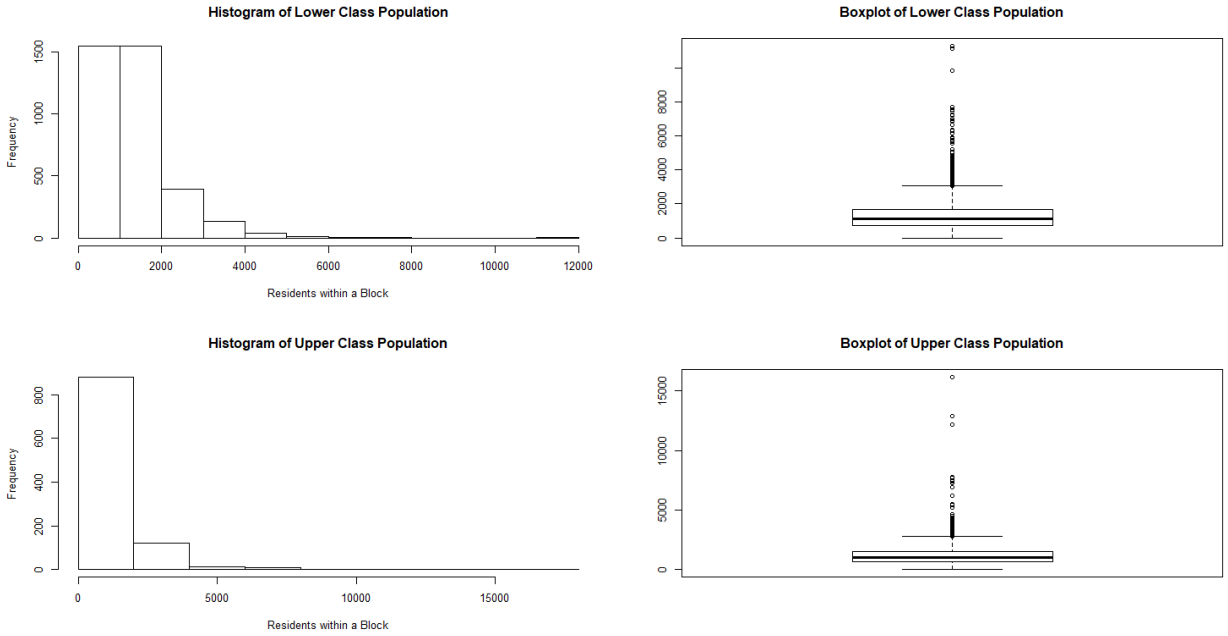

Figure 18

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

Code 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-miss
ing-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-devia
tion/
# 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')
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 buildin
g
# 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)
```

```r
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
# 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')

### Q1
# Refrence: https://bookdown.org/egarpor/NP-UC3M/kde-ii-mult.html
density_estimate <- kde(x = cbind(longitude, latitude))
plot(density_estimate, display = 'persp', ticktype = 'detailed', phi = 25, theta = 10)
```

```r
# Reference: https://www.littlemissdata.com/blog/maps
# map API: AIzaSyBIataTV-gluzSZTnpMSRnlohlwQaNkDDM
ggmap::register_google(key = "AIzaSyBIataTV-gluzSZTnpMSRnlohlwQaNkDDM")
ca_map <- ggmap(get_googlemap(center = c(lon = -119.4179, lat = 36.7783),
                              zoom = 6, scale = 2,
                              maptype ='terrain',
                              color = 'color'))
ca_points <- ca_map + geom_point(mapping = aes(x = longitude, y = latitude),
                                 data = housing, alpha = 0.3, size = 0.5)

# heatmap
# Reference: https://stackoverflow.com/questions/32148564/heatmap-plot-by-value-using-ggmap
ca_heatmap <- ca_map + stat_density2d(
  mapping = aes(x = longitude, y = latitude, fill = ..level.., alpha = ..level..),
  data = housing, geom = 'polygon', size = 0.01, bins = 100) +
  scale_fill_gradient(low = 'red', high = 'green') +
  scale_alpha(range = c(0, 0.3), guide = FALSE)

# Reference: https://stackoverflow.com/questions/32946539/maps-printing-one-by-one-and-not-in-
one-row-using-ggmap-loop-mfrow
grid.arrange(ca_points, ca_heatmap, ncol = 2)

### Q2
indices_958 <- which(housing[,"median_house_value"] == max(median_house_value))
med_val_958 <- housing[indices_958,]

ca_958 <- ca_map + geom_point(mapping = aes(x = longitude, y = latitude),
                              data = med_val_958, alpha = 1, size = 0.5)
ca_958_heatmap <- ca_map + stat_density2d(
  mapping = aes(x = longitude, y = latitude, fill = ..level.., alpha = ..level..),
  data = med_val_958, geom = 'polygon', size = 0.01, bins = 100) +
  scale_fill_gradient(low = 'red', high = 'green') +
  scale_alpha(range = c(0, 0.3), guide = FALSE)

sf_map <- ggmap(get_googlemap(center = c(lon = -122.2913, lat = 37.8272),
                              zoom = 9, scale = 2,
                              maptype ='terrain',
                              color = 'color'))
sf_map2 <- sf_map + geom_point(mapping = aes(x = longitude, y = latitude),
                    data = med_val_958, alpha = 1, size = 0.7, color = 'red')

la_map <- ggmap(get_googlemap(center = c(lon = -118.2437, lat = 34.0522),
                              zoom = 9, scale = 2,
                              maptype ='terrain',
                              color = 'color'))
la_map2 <- la_map + geom_point(mapping = aes(x = longitude, y = latitude),
                    data = med_val_958, alpha = 1, size = 0.7, color = 'red')

grid.arrange(sf_map2, la_map2, ncol = 2)

par(mfrow = c(2,2))
plot(kde(med_val_958$total_rooms), display = 'persp')
plot(kde(med_val_958$median_income), display = 'persp')
plot(kde(med_val_958$housing_median_age), display = 'persp')
barplot(prop.table(table(med_val_958$ocean_proximity))[order(prop.table(table(med_val_958$ocea
n_proximity)))], main = 'Percentage Histogram of Ocean Proxmity')

### Q3
hist(median_income)
low_to_high_inc <- housing[order(housing$median_income),]
```

```r
bottom_18 <- round(nrow(housing) * 0.18)
top_5 <- round(nrow(housing) * 0.05)
lower_class <- low_to_high_inc[1:bottom_18,]
upper_class <- low_to_high_inc[(nrow(housing) - 1 - top_5):nrow(housing),]


ca_map + geom_point(mapping = aes(x = longitude, y = latitude),
                    data = lower_class, alpha = 0.5, size = 0.2, color = 'red') +
  geom_point(mapping = aes(x = longitude, y = latitude),
                    data = upper_class, alpha = 0.5, size = 0.2, color = 'blue')

par(mfrow = c(1,2))
barplot(prop.table(table(lower_class$ocean_proximity))[order(prop.table(table(lower_class$ocea
n_proximity)))])
barplot(prop.table(table(upper_class$ocean_proximity))[order(prop.table(table(upper_class$ocea
n_proximity)))])
prop.table(table(lower_class$ocean_proximity))
prop.table(table(upper_class$ocean_proximity))

par(mfrow = c(2,2))
hist(lower_class$population, main = 'Histogram of Lower Class Population', xlab = 'Residents w
ithin a Block')
boxplot(lower_class$population, main = 'Boxplot of Lower Class Population')
hist(upper_class$population, main = 'Histogram of Upper Class Population', xlab = 'Residents w
ithin a Block')
boxplot(upper_class$population, main = 'Boxplot of Upper Class Population')
median(lower_class$population)
median(upper_class$population)
```