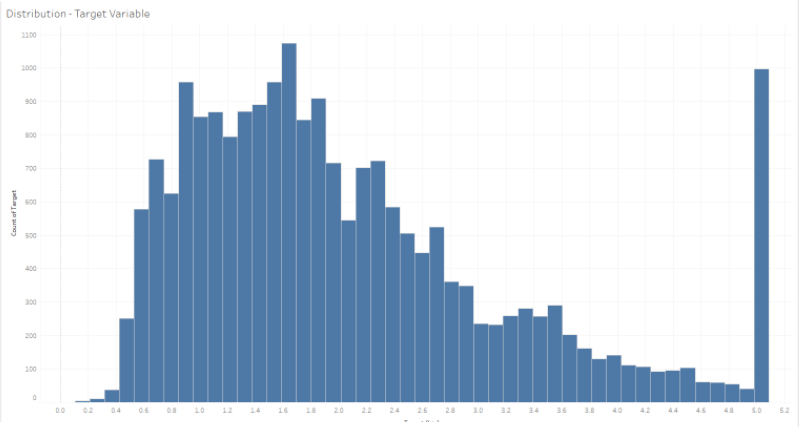
**Data Visualization**

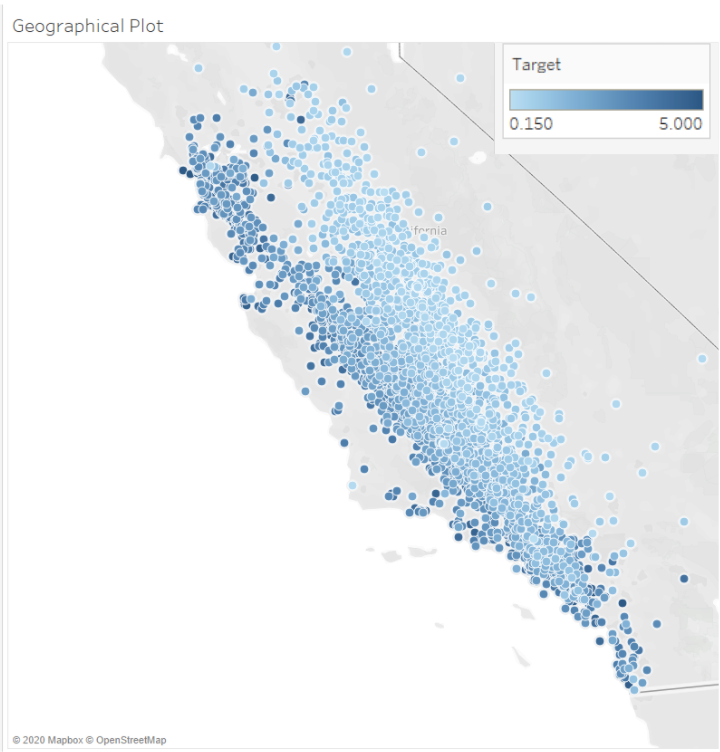
HW 1

09/08/2020

The data set chosen for this project was sources from scikit learn and is titled “California Housing Data” and is primarily used for training linear regression models.

The dataset comprises 9 variables in total. The target variable is the median housing value in California. In addition, there are 8 dependent variables, which include average number of bedrooms, average occupancy, average number of rooms, house age, latitude, longitude, median income and population. All are continuous variables. The dataset did not include any missing values.

The purpose of the tableau workbook was to conduct exploratory data analysis on this dataset. The first plot in the workbook is the distribution of the target variable. We can see from the plot that it is right skewed and that there appears to be some anomalies around bin 5 where there is a spike in the count.

The second plot is a geographical plot of our target variable using the longitude and latitude for spatial coordinates. Each circle represents the coordinates of one target observation and the color scale is indicative of the value. From this plot we can see that there the median property value tends to increase as we closer to the coastline, which is to be expected.

The next series of images are scatter plots that compare our target and feature values. Here we are looking to see if there is a positive correlation between our target variable (y-axis) and our feature variables (x-axis). A visual inspection of these plots, with the addition of a regression line, appears to indicate that there a positive relationship between the target variable and median income. This seems to be logical and what one would expect.

