The next work is based on a dataframe from houses of Tucson Arizona that was previously clean, among the variables can be find coordinates, sold\_price, sqrt feet, bedrooms, bathrooms, fireplaces, year of built, etc.

Since this dataframe misses a linear tendency, the prediction started with a classification using the haversine function (convert the coordinates into a number) applying a lambda in order to take the variables directly from the dataframe as ‘x’ and a new variable (sold price/ sqrt feet) as ‘y’, after check the histogram, it was expected to have , it was found 5 categories ( 5 was add to 4 since they were few observations), those categories follow a 150 range, since the accumulation of observations are in majority center in 200. The objective of these variables was to add sensibility to the model for the change of price according to the quantity and place where the property is found.

Having the categories assigned, it was decided to use OLS for creating the regressions, the x variables were bathrooms, bedrooms and sqrt feet since they were the ones with highest correlation. The R2 goes from 73% to 90%, the saturation of category 2 gives us a detailed model while category 1 goes to 100 observations.

Originally it was intended to predict how much the price for airbnb would be, but it was missing a critical variable, the score of reviews. The model broke at the predictions of price per day, since the expected among was 200-500 per night and the results were like estimations of the whole sale of the house.