The regression main project consisted in developing a model that could predict the price of a house based on specific features provided from the dataset. There were some side-tasks like answering some specific python questions in Jup-notebook, creating a notebook with a data analysis and more.

To develop the model, we had to understand what data we worked with, so we decided to plot different graphs:

* Main graph was a global position one (latitude-longitude) where we could see what’s the most common placing for all the properties on the globe.
* Different graphics making comparisons of price with different parameters like number of bedrooms, grading, conditions, area of living and many more, basically using most of the columns that were available to us in the raw data excel and csv docs.
* Checked correlation between few variables

Gráfico

Descripción generada automáticamenteAfter all that analysis we proceeded to start with the scaling and machine learning process: imported all the needed “sklearn” libraries, proceeded with “train” and “test” coding for the regression. Considering all the data we had, which was quite a lot, we proceeded with a testing process of 10% and train of 90%. After reg.fit, we managed to get a regressor score of 70.5% which was not enough to comply with a decent minimum, so we continued with “GradientBoostingRegressor” from the ensemble library (sklearn), managing a score of 86.3%, which could be considered valid for the machine learning process.

Concluding the model testing and all necessary data analysis we concluded that, even though price can be sorted by absolutely every single condition (columns), there were few that made quite an interesting difference, specially for properties over 650k:

All properties within a longitude of -122.2º and latitude of 47.6º *(which we also decided to search, seems that all the properties are located close to Seattle, WA)*, tend to be way more expensive than in any other location, zip codes around 98040 and 98000 have high priced housing, so we could assume it’s located within those specific coordinates. Also, grading was very directly proportional to pricing in some way since most expensive houses also had the highest grades.

An interesting fact is that most of the expensive houses have an estimated condition of 3.