A Data Science Perspective

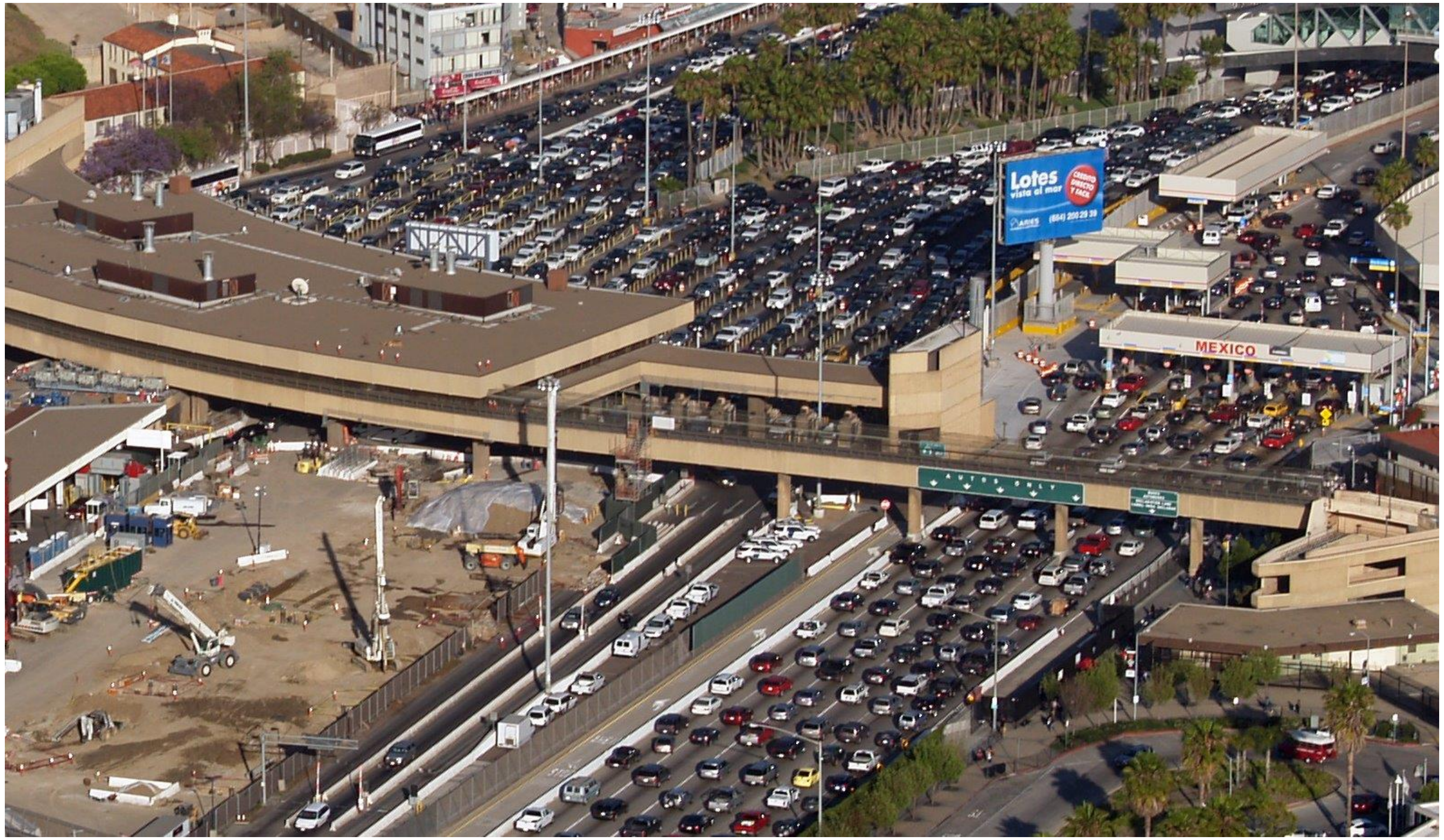
92173

Scope: Visualization and Machine Learning

We will start by doing a real estate analysis of the 92173 Zip code.

This is a Zip code very dear to me as I have lived here for more than 10 years.

Code will be provided on a GitHub link.



Between Two Worlds

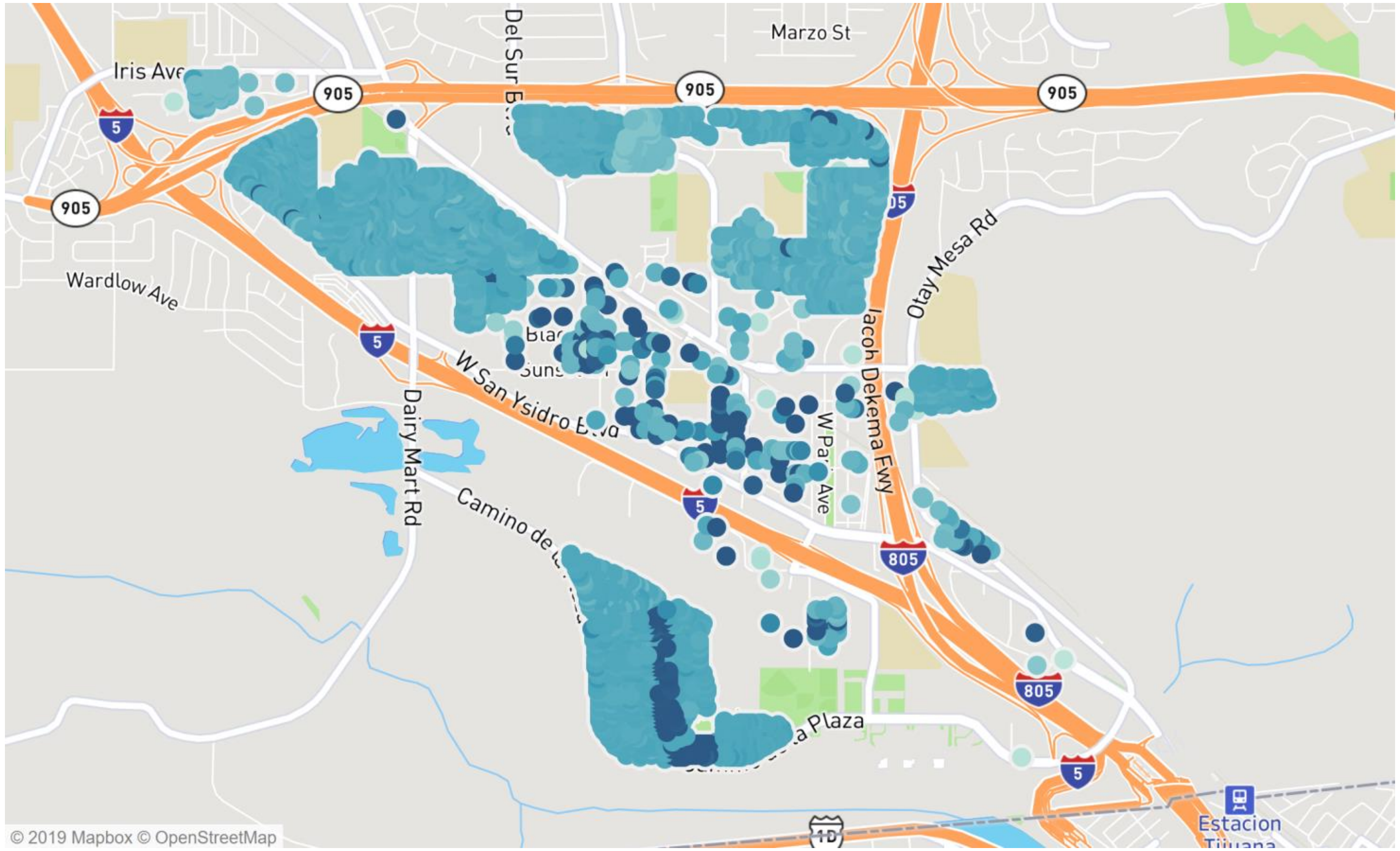
- The city of San Ysidro has been around since the 1900’s.
- Serves as a gate between Mexico and the United States.
- Also a rich community with a 93.8% Hispanic demographic makeup.

Source:https://statisticalatlas.com/neighborhood/California/San-Diego/San-Ysidro/Race-and-Ethnicity

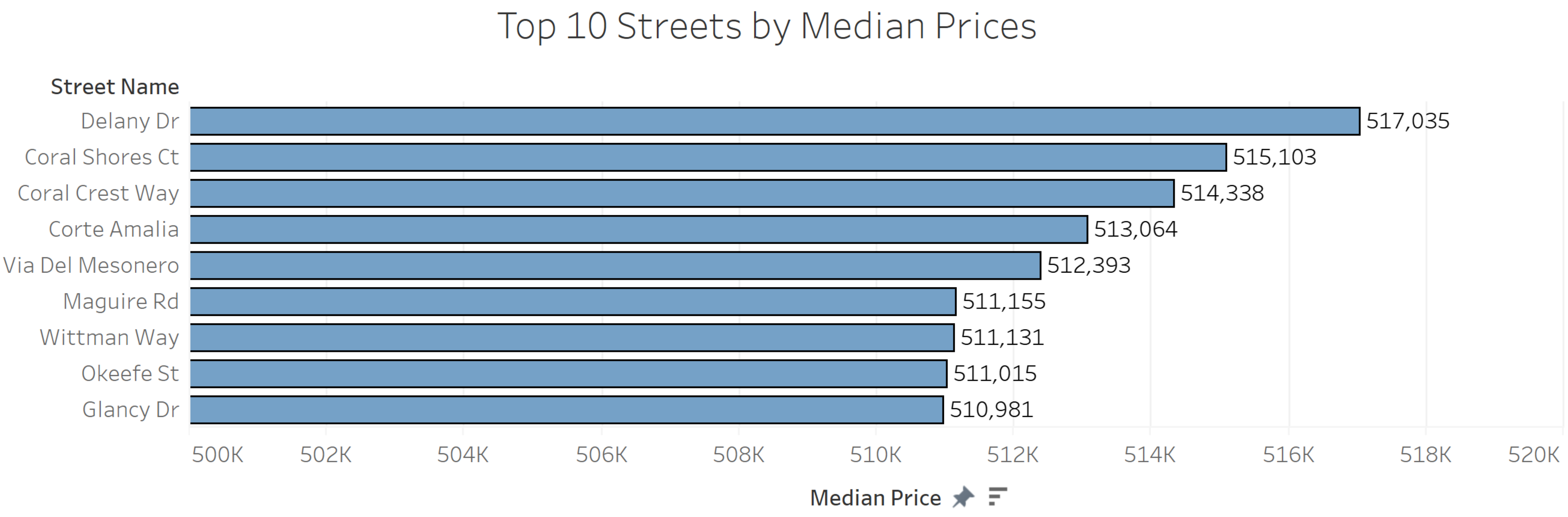
Initial Look at Housing

- After connecting to the Zillow Api I managed to scrap data for 2415 residencies.
- They are color coded by price, the darker the color the higher the price.
- Full interactive map can be found here:
    - https://qrgo.page.link/tGB89

San Ysidro House Map by Prices

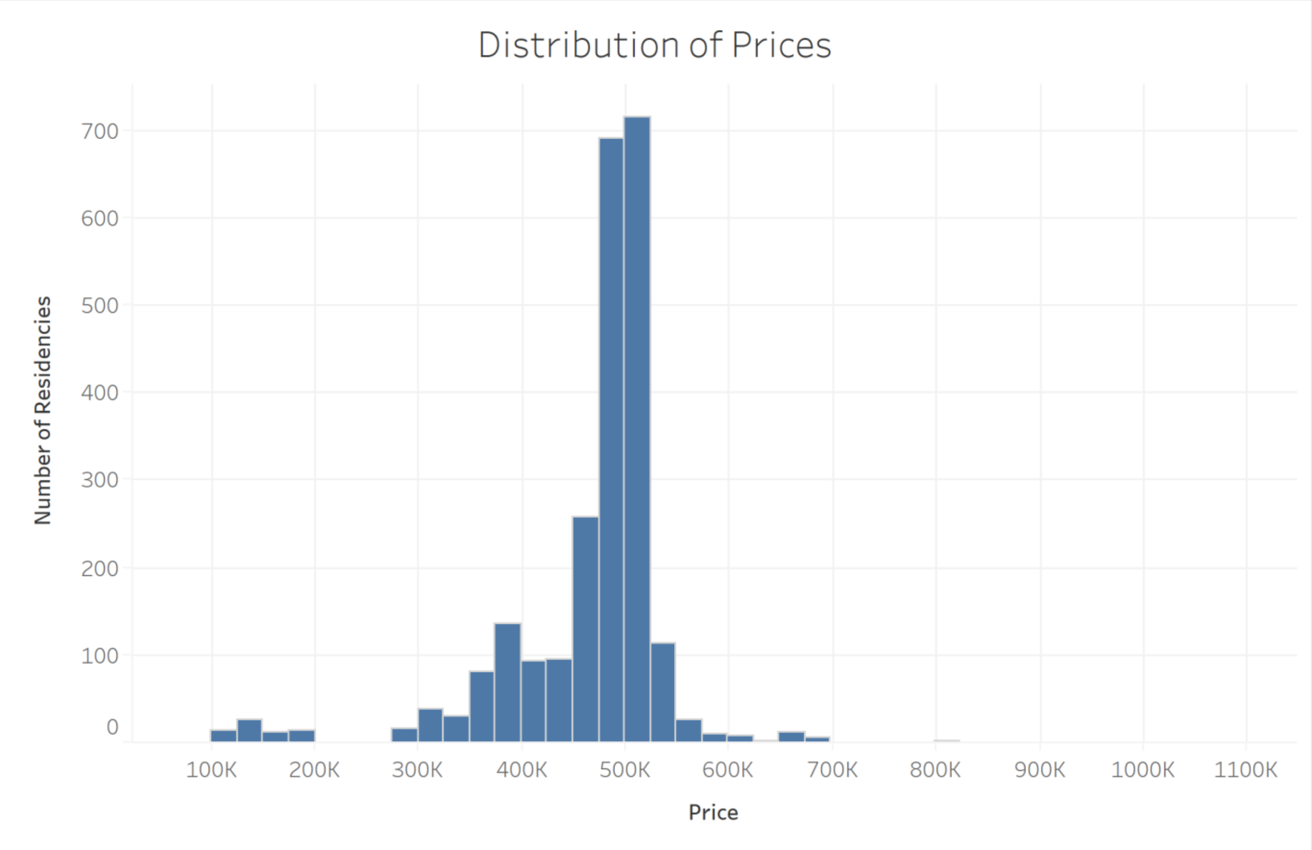


Most Valued Streets



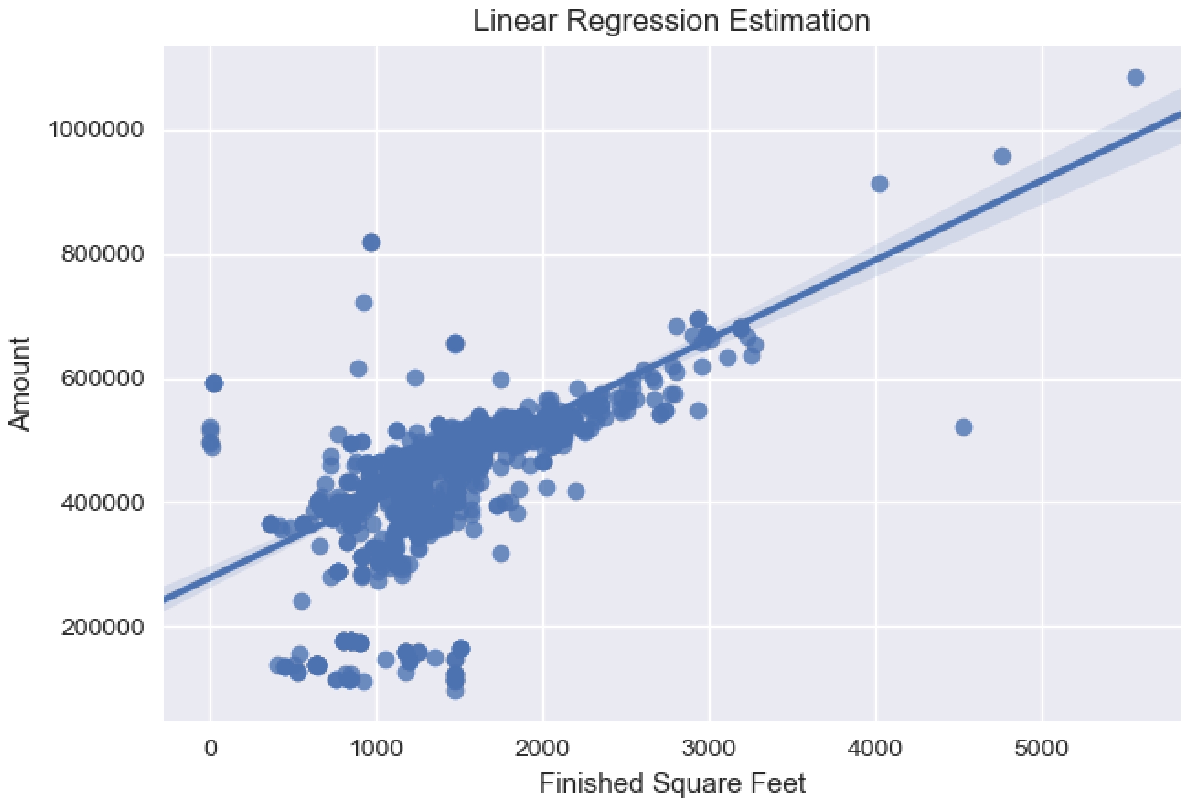
Distribution

- Constructing a histogram we can take away the following.
  - Most houses are between 450K – 550K USD.
  - There is a small cluster of residencies from 100K to 200K USD.
  - This is follows a normal distribution.



Regression Models (Price)

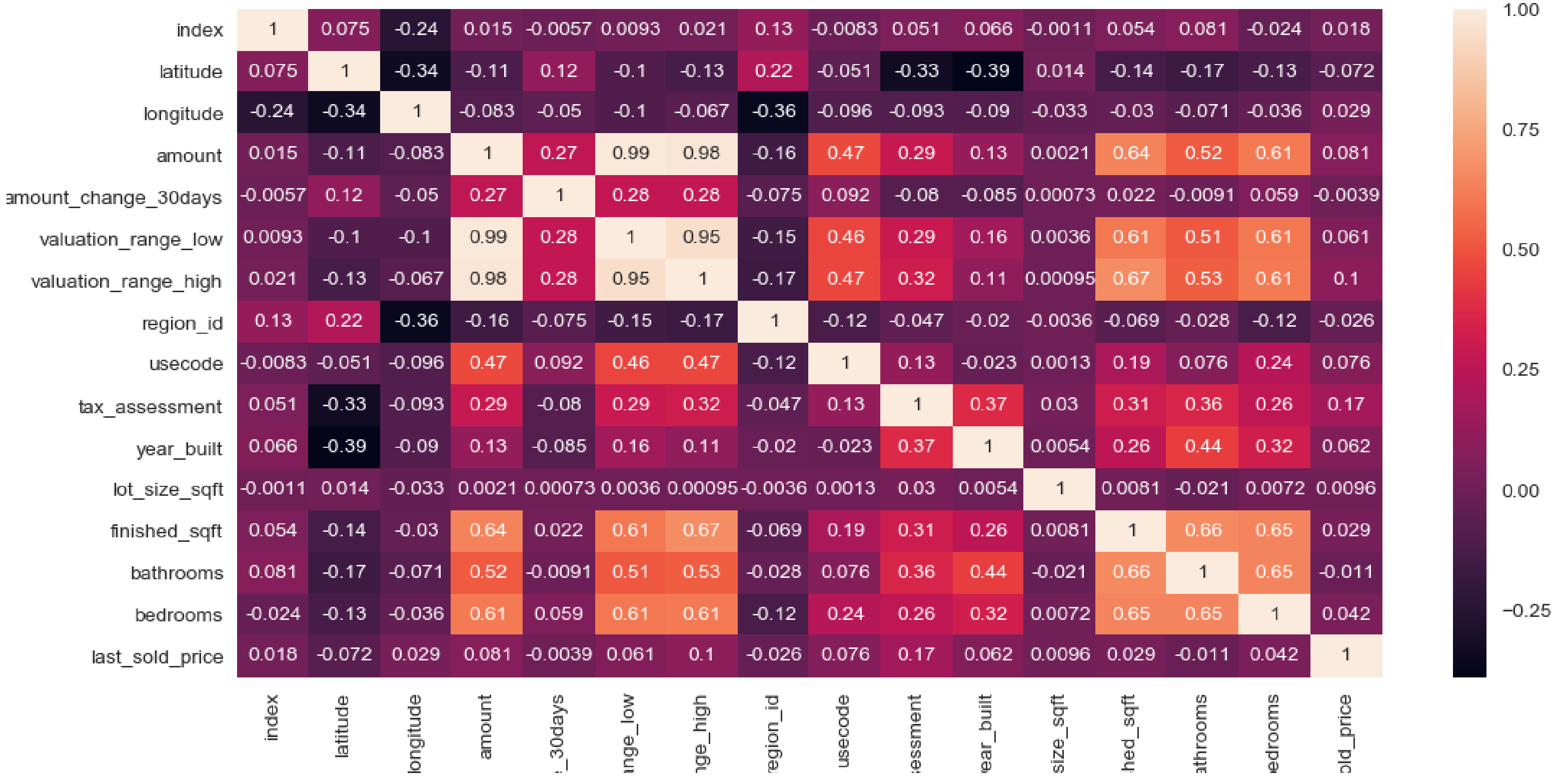
- This type of continuous data on Prices requires regression techniques to make estimates.
- On the right we can see a method known as Linear Regression minimized by OLS.
- We draw the line of best fit and predict values based on that.



Feature Selection and Correlation

- Before we perform regression techniques for prediction, we have to pick which features are most important.
- A good step is to draw a correlation matrix and choose
- We chose:
  - Usecode
  - Finished_sqft
  - Bathrooms
  - Bedrooms

Correlation Matrix

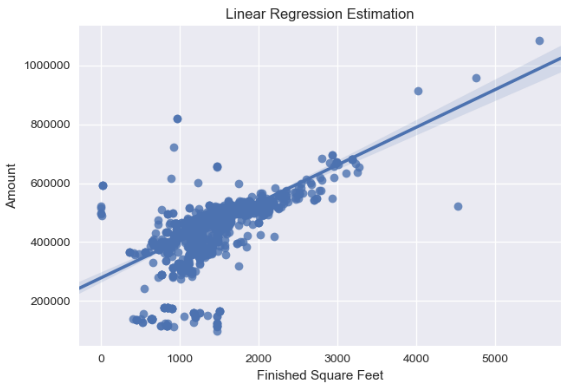


	index	latitude	longitude	amount	_30days	ange_low	nge_high	region_id	usecode	essment	ear_built	size_sqft	hed_sqft	athrooms	edrooms	old_price
index	1	0.075	-0.24	0.015	-0.0057	0.0093	0.021	0.13	-0.0083	0.051	0.066	-0.0011	0.054	0.081	-0.024	0.018
latitude	0.075	1	-0.34	-0.11	0.12	-0.1	-0.13	0.22	-0.051	-0.33	-0.39	0.014	-0.14	-0.17	-0.13	-0.072
longitude	-0.24	-0.34	1	-0.083	-0.05	-0.1	-0.067	-0.36	-0.096	-0.093	-0.09	-0.033	-0.03	-0.071	-0.036	0.029
amount	0.015	-0.11	-0.083	1	0.27	0.99	0.98	-0.16	0.47	0.29	0.13	0.0021	0.64	0.52	0.61	0.081
amount_change_30days	-0.0057	0.12	-0.05	0.27	1	0.28	0.28	-0.075	0.092	-0.08	-0.085	0.00073	0.022	-0.0091	0.059	-0.0039
valuation_range_low	0.0093	-0.1	-0.1	0.99	0.28	1	0.95	-0.15	0.46	0.29	0.16	0.0036	0.61	0.51	0.61	0.061
valuation_range_high	0.021	-0.13	-0.067	0.98	0.28	0.95	1	-0.17	0.47	0.32	0.11	0.00095	0.67	0.53	0.61	0.1
region_id	0.13	0.22	-0.36	-0.16	-0.075	-0.15	-0.17	1	-0.12	-0.047	-0.02	-0.0036	-0.069	-0.028	-0.12	-0.026
usecode	-0.0083	-0.051	-0.096	0.47	0.092	0.46	0.47	-0.12	1	0.13	-0.023	0.0013	0.19	0.076	0.24	0.076
tax_assessment	0.051	-0.33	-0.093	0.29	-0.08	0.29	0.32	-0.047	0.13	1	0.37	0.03	0.31	0.36	0.26	0.17
year_built	0.066	-0.39	-0.09	0.13	-0.085	0.16	0.11	-0.02	-0.023	0.37	1	0.0054	0.26	0.44	0.32	0.062
lot_size_sqft	-0.0011	0.014	-0.033	0.0021	0.00073	0.0036	0.00095	-0.0036	0.0013	0.03	0.0054	1	0.0081	-0.021	0.0072	0.0096
finished_sqft	0.054	-0.14	-0.03	0.64	0.022	0.61	0.67	-0.069	0.19	0.31	0.26	0.0081	1	0.66	0.65	0.029
bathrooms	0.081	-0.17	-0.071	0.52	-0.0091	0.51	0.53	-0.028	0.076	0.36	0.44	-0.021	0.66	1	0.65	-0.011
bedrooms	-0.024	-0.13	-0.036	0.61	0.059	0.61	0.61	-0.12	0.24	0.26	0.32	0.0072	0.65	0.65	1	0.042
last_sold_price	0.018	-0.072	0.029	0.081	-0.0039	0.061	0.1	-0.026	0.076	0.17	0.062	0.0096	0.029	-0.011	0.042	1

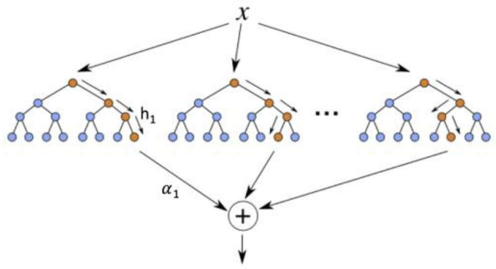
Machine Learning Selection

- We want to now guess and be able to predict the log(price) of a home with our selected features.
- We Will use three different models and see how they each compare.
- Three models:
  - Linear Regression
  - Random Forest
  - XGBoost

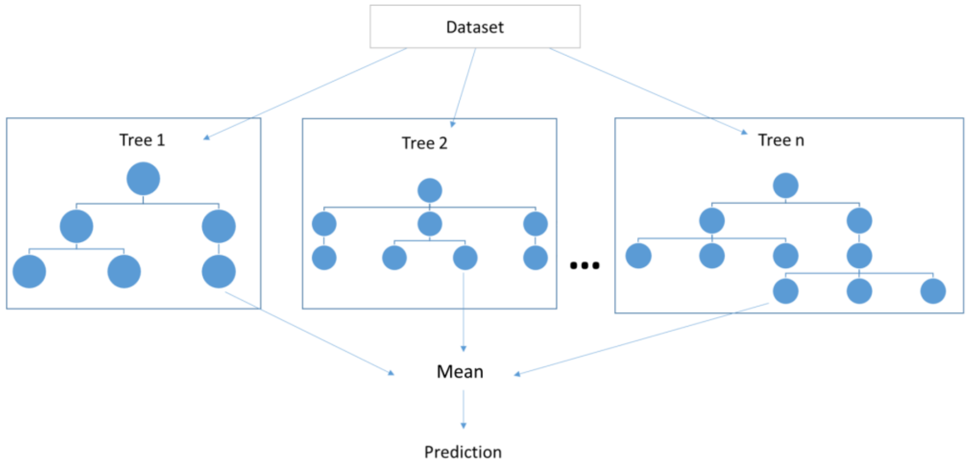
Linear Regression



XGBoost

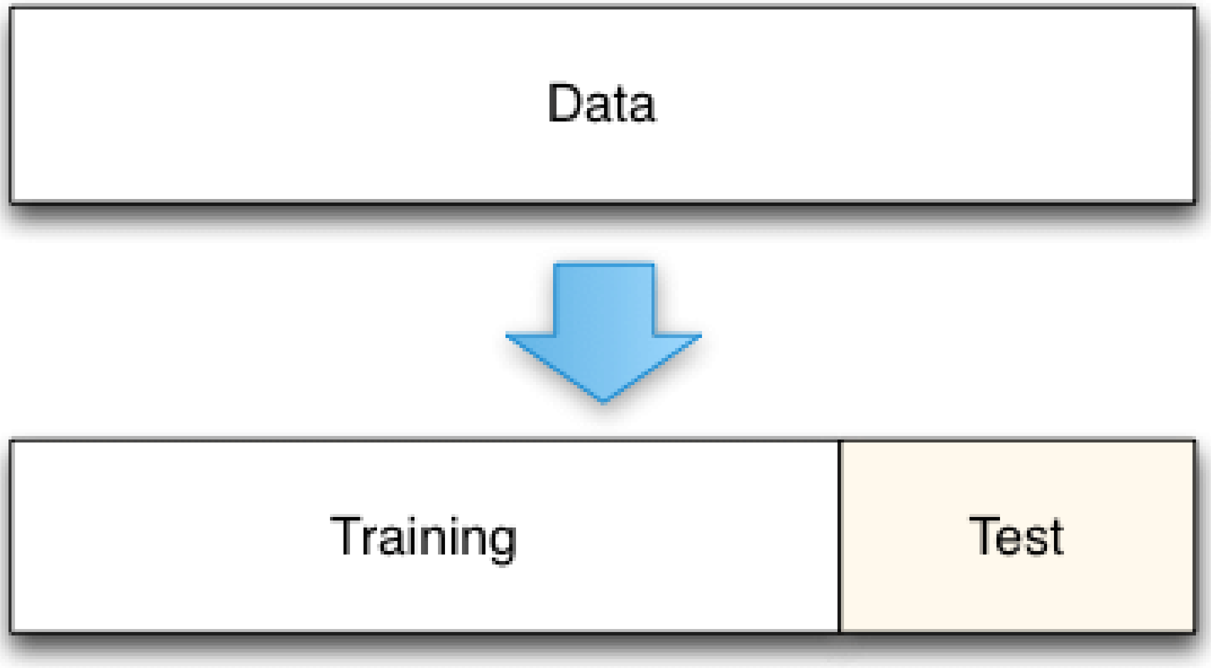


Random Forest



Training and Testing

- We split our data into two data sets training and testing, 70% and 30% respectively.
- This is done to test our machine learning model on observations it theoretically has not seen before.



Machine Learning Results

After our results we can see that the lowest absolute error comes from a random forest regression of 0.090. We will interpret this in the next slide

Target Variable	**Log(Amount)**		
Model	Linear Regression	**Random Forest**	XGBoost
MAE	0.105	**0.090**	0.102

Interpretation

Our best performing model was our Random Forest Regressor which scored a Mean Absolute Error of about 0.09 being the smallest error.

This mean that our model was approximately 9% off from the geometric mean.

This is an alright score and we theoretically should have scored higher ,but there are some issues within our data gathering process.

Analysis Problems

To improve further models we can think about our data gathering. The Zillow API was constantly missing observations of mobile homes and apartment complexes.

These two types make up a great deal of the San Ysidro community and not including them raised the overall mean of the zip code.

There was a systematic bias in our data collection which in turn is reflected in our models.