Basically, for the exploration phase, explore however we want to. Free form.

look at what we’ve done in the past, use functions,

Create at one t-test and one chi sqare test. Basically: EXPLORE THE DATA, and ignore the lesson:

1. Write function that outputs (using the numeric data?) Pairplot, with heatmap in subplots
2. We did this in the exploration of the earlier methodology lessons. Copy these steps.

Use log error for the t-test (since it’s normally distributed) as the Y and any categorical variable as the x, … we did this in class, by slicing into a field (ex: “more than 5 brs vs 3 br, etc”)

Look at the many fields (features) and try to remove as many as possible, ask yourself, “is this field really going to be a big driver? Can I just delete it?”

**Project**

Now move on to exploring with clustering

Decide how many clusters.

This is what Maggie did in class yesterday.

Using Kmeans, figure out what k number is, use elbow method, look for the sharp curve, it’s in the lesson,

Plot, using k as color

It will cluster based on the difference in log error, based on certain features like bedrooms, baths, etc.

Cluster independent variables, like location, size, and then match with log error for example

This will help to decide which features are more important

Test the significance of clusters, using statistical testing methods to determine whether the clusters I created are significant in terms of their relationship to logerror (the dependent variable).

Don’t get caught up in the details. PRESS FORWARD and don’t get caught up in making the perfect cluster. Do one or two and move forward.

If I have time to come back later to do more, great. If not, great. KEEPR PRESSING FORWARD.

NOW REMOVE VARIABLES NOT NEEDED WANTED USEDFUL OR REDUNDANT.

DON’T SPLIT UNTIL READY TO MODEL!

**Train test split and modeling:**

Now:

Train 3 models, likely one will be linear regaression, but try a couple others also.

Use absolute log error if doing a linear regression

Copy this from the previous Zillow

Can create a separate field for the absolute log error.