**Zillow**

Create a file named prepare\_zillow.py that works with the zillow data and contains functions that do the following:

* ~~include only single unit properties (e.g. no duplexes, no land/lot, ...) For some properties, you will need to use multiple fields to~~**~~estimate~~**~~whether it is a single unit property.~~
* takes in a dataframe and a list of columns names and returns the dataframe with the datatypes of those columns changed to a non-numeric type
  + use this function to appropriately transform any numeric columns that should not be treated as numbers
* ~~accepts the unprepared zillow data frame and applies all the transformations above.~~

**Handle Missing Values**

1. Write or use a previously written function to return the total missing values and the percent missing values by column. ***- I have this in telco or previous zillow***
2. Write or use a previously written function to return the total missing values and the percent missing values by row. . ***- I have this in telco or previous zillow***
3. Write a function that will take a dataframe and list of column names as input and return the dataframe with the null values in those columns replace by 0. **-I think I have this already too**
4. ~~Impute the values in land square feet.~~

~~For land square feet, the goal is to impute the missing values by creating a linear model where landtaxvaluedollarcnt is the x variable and the output/y-variable is the estimated land square feet. We'll then use this model to make predictions and fill in the missing values.~~

~~Write a function that accepts the zillow data frame & returns the data frame with the missing values filled in.~~

1. Create a function that fills missing values with 0s. Explore the data and decide which columns it makes sense to apply this transformation to.  **– this is in telco or previous zillow**
2. Run the first function that returns missing value totals by column: Does the attribute have enough info (i.e. enough non-null values) to be useful? Choose your cutoff and remove columns where there is not enough information available. Document your cutoff and your reasoning. **-- This is done from lesson**
3. Run the function that returns missing values by row: Does the observation have enough info to use in our sample? Choose your cutoff and remove rows where there is not enough information available. Document your cutoff and your reasoning**. -- This is done from lesson**
4. ~~Of the remaining missing values, can they be imputed or otherwise estimated?~~
   * ~~Impute those that can be imputed with the method you feel best fits the attribute.~~
   * ~~Decide whether to remove the rows or columns of any that cannot be reasonably imputed.~~
   * ~~Document your reasons for the decisions on how to handle each of those.~~

**Handle Outliers - THIS IS IN THE PREPARE\_MALL.PY FILE**

1. Write a function that accepts a series (i.e. one column from a data frame) and summarizes how many outliers are in the series. This function should accept a second parameter that determines how outliers are detected, with the ability to detect outliers in 3 ways:
   * **This is done already, code/methodology from the lesson.**
   * Using the IQR
   * Using standard deviations
   * Based on whether the observation is in the top or bottom 1%.
2. Use your function defined above to identify columns where you should handle the outliers.  **THIS IS THE OUTLIER PEEK IN THE PREPARE.MALL.PY file**
3. Write a function that accepts the zillow data frame and removes the outliers. You should make a decision and document how you will remove outliers.
4. Is there erroneous data you have found that you need to remove or repair? If so, take action.
5. Are there outliers you want to "squeeze in" to a max value? (e.g. all bathrooms > 6 => bathrooms = 6). If so, make those changes.