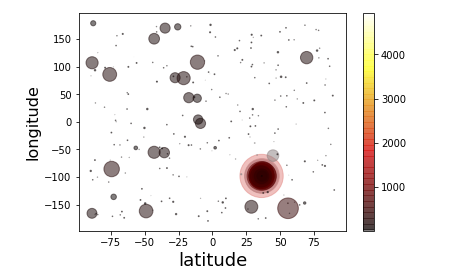
1. **Problem Statement**: To evaluate a KNN algorithm to estimate housing prices using a housing dataset with the features of location, time and price. The model must ensure that it prevents time leakage i.e. a neighboring houses price should only be considered in the model if its closing price date is less than the house we are trying to predict the price of.
2. **Goal**:
   1. To minimize the Median Relative Absolute Error with k = 4
   2. To observe and take care of spatial/temporal trends
   3. To improve on the model if time permits or provide recommendations
   4. To provide recommendations for productionizing the model
3. **Timeline**:
   1. Research and preliminary data analysis for the project: 20 minutes
   2. Coding and optimizing code: 2 hours
   3. Ran code and wrote out this report simultaneously: 30 minutes
   4. Checked for spatial and temporal trends with MRAE: 10 minutes
4. **Questions:**

* Using the dataset provided, please build a *k*-NN model for *k = 4* that avoids time leakage (details below).
  + The code for the model can be found in the file LeakyKNN.py.
  + To run it simply run python LeakyKNN.py. To change number of rows that are being processed change the value of N in the last line of the python file.
  + Code was run for 30 minutes which allowed for n=16000 rows to be processed.
* What is the performance of the model measured in Median Relative Absolute Error?
  + The MRAE for my model is: 0.21120559463943298
* What would be an appropriate methodology to determine the optimal *k*?
  + The optimal methodology would be using K-fold (K is unrelated to optimal k) cross validation. We can then try different values of k and plot that versus the MRAE and select the model with best MRAE and lowest k. Using this technique the optimum k can be found.
  + Consulting with experts of the housing market field may provide insight into selecting the optimal k.
* Do you notice any spatial or temporal trends in error?
  + Due to time constraints, I only investigated this very briefly. With more time I would like to examine temporal trends to a greater detail such as finding any seasonal trends or comparing the rates of error for different years and/or months.
  + Spatially in the first 1000 houses larger errors were found in the range of latitude - (-150,-100) and longitude (25,50). I would consult with experts and examine the data in these regions more closely and see if improvements can be made to the model or if some spurious rows need to be dropped to improve the predictions.
  + Visualized below with size and color assigned using RAE.



* How would you improve this model?
  + Right now the model only makes use of the price and distance – it is not using the time. I would add a weight for the time a house was sold too, so that houses which are sold closer to the closing date of the house in consideration could be given more weightage. I would then verify if this hypothesis provides better predictions.
  + I would request other data features such as number of bedrooms, square footage, parking and furnishing and make use of them in improving the model. I could then implement a different method such as regression and see how the performance of the model improves using these various features.
* How would you productionize this model?
  + The current model is not optimized for speed and runs with O(N2) time complexity. Hence the first step to ensure this model could be productionized is to improve the speed of the algorithm.
  + The KNN algorithm is highly parallelizable as each iteration of the loop does not depend on the other iterations. So the brute force KNN that we are using could be easily partitioned as no communication or message passing would be required.
  + Alternatively, I could look into other techniques such as regression or even neural networks where once a model is trained we no longer need to keep performing time consuming calculations hence having the model released would be more feasible.