# Capstone Project 1 Machine Learning
# Predicting Wind Farms Impact on Home Values

### Current Interpretation of Data:

The results in the Milestone Report for linear regression and statistical analysis (slides 8 – 12 at this link), show that *population density* appears to have a positive correlation with home values. However, it's also interesting to note that the level of correlation between zipcodes with windfarms versus without windfarms is negligible at only 2% (23% vs 25%) difference.

Regarding median income, correlation to home values is much lower for zipcodes with a windfarm than zipcodes without a windfarm (19% versus 41%). A potential reason for this is that windfarms may tend toward more rural areas and in rural areas median income may be less of a factor on home values.

In the results to date the impact on home values of having a windfarm in close proximity does not appear to be statistically significant (a $360,000 mean for homes in zipcodes w/windfarms versus $400,000 for homes in zip codes without windfarms).

To this point the objective of this project has been a regression algorithm that can be used to predict the change in value of a home in close proximity to windmills. However, if the above results receive further confirmation in machine learning the objective will be modified to "a regression algorithm that can be used to predict the change in value of a home ~~in close proximity to windmills~~". In other words, wind farms may very well be demonstrated to not impact home values.

Rather than immediately start running algorithms the approach for machine learning will be to...

1. *Add More Features:* review again the features listed in the zipcode database and consider if there are additional features to bring into this study
2. *Expand Data Used & Quality of Data Used:* review the data used for statistical analysis and identify changes that could improve the quality of the data used for machine learning
3. *Algorithms:* apply the following algorithms to the data to see if a model > 80% accuracy can be created:
    a. Decision Tree
    b. Random Forest
    c. Ridge Regression
    d. Lasso Regression

### Results for Adding More Features:

A review of the zipcodes database identified "level of education attained" as an additional feature to consider for the model. The feature is called "Percent_Higher_Education" in the input data and identifies the percent of the population with an education level of at least an Associates Degree versus an education level of a high school diploma or no diploma. (formula: $\frac{\text{Associates + BS or BA + Masters + Doctorate}}{\text{(No Diploma + HS Diploma + Associates + BS or BA + Masters + Doctorate)}}$ )

### Results for Expand Data Used and Quality of Data Used:

Potential reasons for the lower than expected prediction value of population median and median income are that statistical analysis...

1. did not look at enough data (only the year 2017's housing data was used)
2. the data cleaning may have skewed the results (the "bfill" method was used to address null values)

To rule these out as reasons for poor prediction, machine learning will...

1. span the years 2009-2017

2. break the zip codes into a series of datasets based on quality of housing data available. For example, a zip code with home values from 2009 – 2017 will be included in all datasets. A zip code with home value data starting in 2014 will be included in the 2014, 2015, etc... datasets, but not in the 2009 – 2013 dataset. Taking this approach removes the need to use data cleanup method, such as bfill.

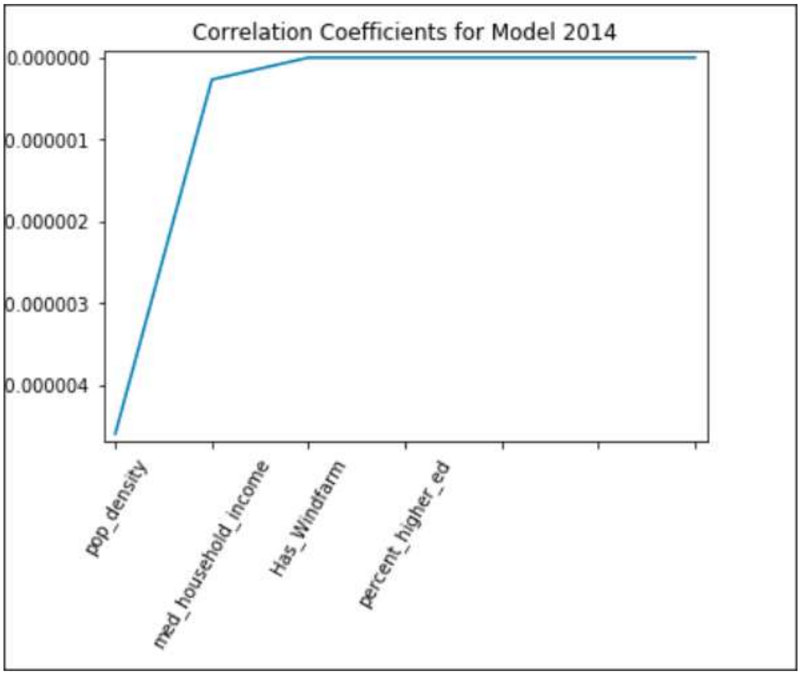
### Results for Algorithms:

For each algorithm (Decision Tree, Random Forest, Ridge, and Lasso) three datasets (2009 – 2017, 2014 – 2017, and 2009 – 2017 without home value features) were used.

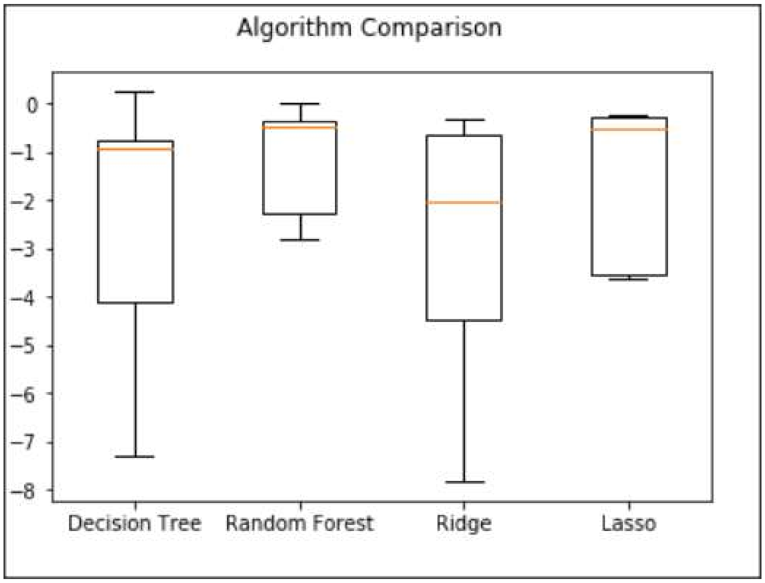
*Some key conclusions are...*

- The most significant predictor is population density, followed closely by median income
- The presence of a windfarm does not appear to have an impact on home values in the surrounding area
- The number of years of home value data (each year being treated as a feature) has some impact on the model, but as previously noted population density and median income's impact is more significant.

  The chart below derived from the correlation coefficients of the Lasso algorithm supports the above statements...



Each algorithm was analyzed individually and was also included in a comparison via the use of a box chart plot...



*There are several points of interest to learn from the box chart...*

- All the algorithms produce a similar top, but Random Forest and Lasso appear to be the best models to use for further analysis. They have the highest top and a much tighter range than Ridge or Decision Tree.
- Random Forest's plot is short and Lasso has very short whiskers indicating a high level of agreement on home value predictions across zip codes.
- Height of the Decision Tree and Ridge plots are similar, and are much greater than Random Forest and Lasso. This lends further support to not continuing with the Decision Tree and Ridge algorithms.
- The medians of Decision Tree, Random Forest, and Lasso are very similar, but the size of their box plot varies, showing a much higher variance in the Decision Tree results.

**Links to Machine Learning Jupyter Notebooks on GitHub:**

Machine Learning6 - Prepare data for Machine Learning

Machine Learning6a - Decision Tree Algorithm

Machine Learning6b - Random Forest Algorithm

Machine Learning6c - Ridge Regression Algorithm

Machine Learning6d - Lasso Algorithm

Machine Learning6e - Compare Algorithms