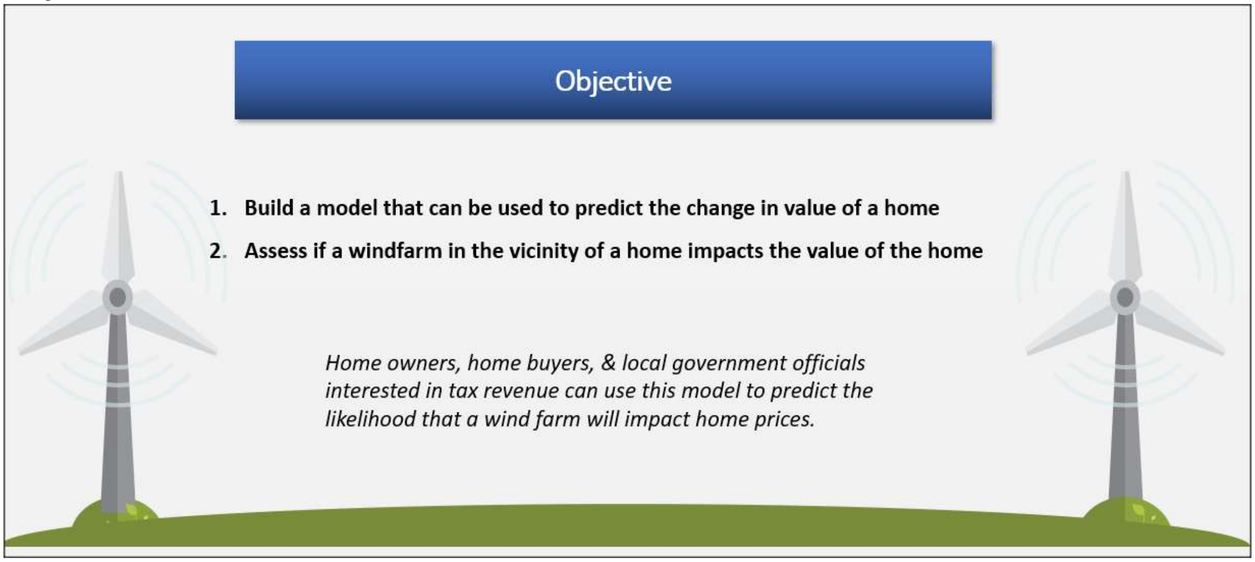
Capstone Project 1 Final Report
Predicting Wind Farms Impact on Home Values

Contents

Capstone Project 1 Final Report
Predicting Wind Farms Impact on Home Values

Objective:



Problem:

There are many factors that influence a home's value such as the region's economy, condition of the home, proximity to shopping & other attractions, school district rating, etc…, but could a wind farm be a greater factor on home values than one, or even a combination of several of these factors.

Imagine you just opened your favorite news feed and the local news is that a wind farm is coming to your area. As a homeowner, should you be concerned about your home's value? As a town board member should you raise concerns about property tax revenue (due to a decrease in home values) at the next meeting?

We propose to use data from various sources on wind farms and home values to build a model for predicting home values and answer the question above regarding do windfarms impact home values. The model will calculate the year-over-year changes in home value for zip codes with a windfarm and compare it to the year over year home value changes zip codes that do not have a windfarm.

Client:

Home owners, prospective home buyers, and local government officials interested in tax revenue can use this model to predict future home value and the likelihood that a wind farm will impact home prices.

Dataset:

The wind farm data is obtained from the US Wind Turbine database. This data contains numerous information about each wind mill in each wind farm; of interest for this study is the windmill Id, project id, and location (latitude, longitude, zip code, and state).

In addition, dataset/s containing information about home values is needed for this study. This information is obtained from Zillow housing data.

Dataset parameters:

- The dataset timeframe for wind farm data is farms constructed in the years 2009 – 2017.
- Zipcodes with windfarms and zip codes without windfarms 25 miles away from the zip code with a windfarm were evaluated

Approach:

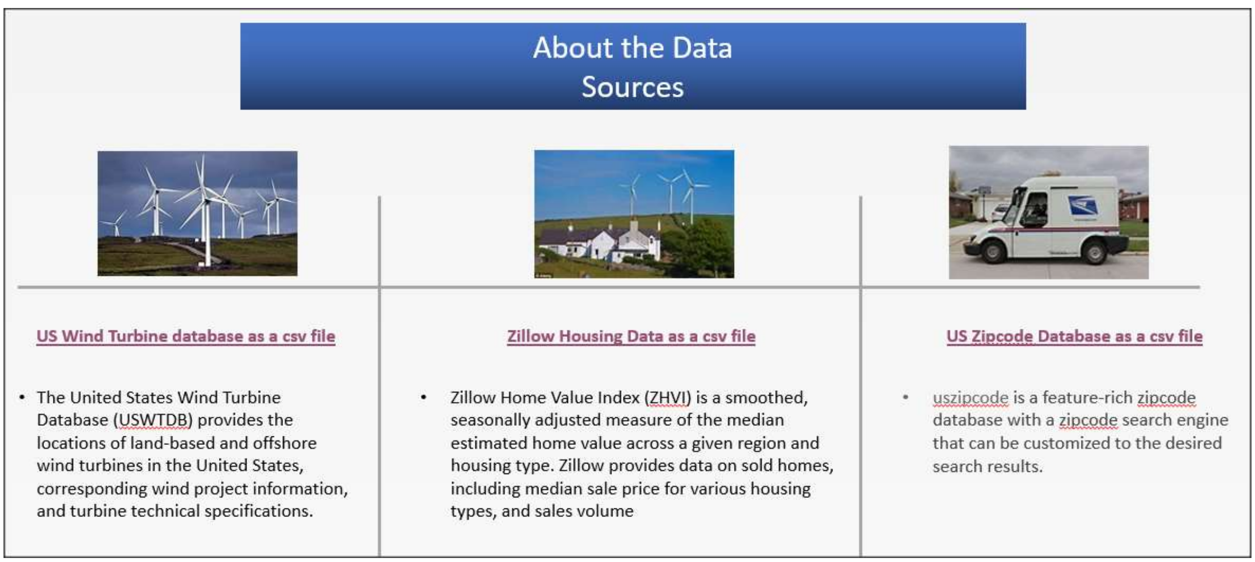
Since the objective is to assess if percentage change in home values vary between zipcodes with farms and zipcodes without, several algorithm will be tried to build a model with predictive ability. The model will provide a prediction of the percentage change in home value in future years.

The predictor variables will be…

- windfarm or no windfarm in a given zip code
- houses in a given zipcode (location, percentage change in value in prior years, median income, population density, and education level)

About the Data

Prior to using the data a review of the US Wind Turbine and Zillow Housing data was completed. The image below summarizes the data source and approach used to obtain data...



Data Cleaning Steps:

Overall the data was very clean and required minimal clean-up work. The Zillow housing data proved to be less complete than initially thought (more on this later in this paper) and the US Zipcode database proved useful to obtain zip code information not contained in the Zillow housing data.

Here are the steps taken to clean each data source...

US Wind Turbine database (link)

Downloaded the US Wind Turbine database as a csv file
Downloaded and reviewed the codebook, which provides a description of each column in the file
Flagged 10 of 24 total fields (columns of data) that are of value for this study on wind farms

Loaded the csv file
Created a new csv file of the 10 desired fields

Completed a % of clean data assessment on each field. Using the pandas dataframe ".info" command, the data comes back as very clean…

Visual analysis showed that empty string fields in this file are populated with the word, "missing". However, an analysis of the data in the columns needed came back with no missing data.

This study relies on each windfarm having a zip code. Windfarms with geocodes that failed the zip code lookup were assigned a value of '99999'.

Zillow Housing data (link)
Downloaded the housing data from Zillow

Noted constraint of only median-sale price data is available (desired average sales price as well)

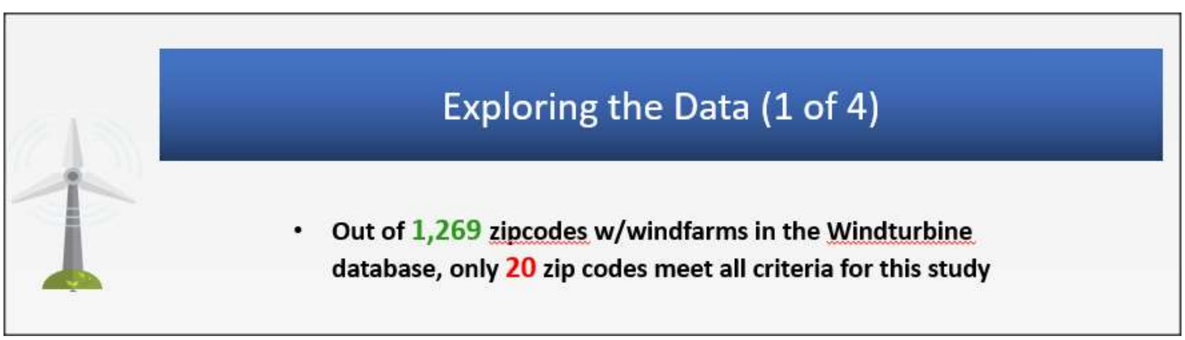
Data initially looked clean, but that was because missing data was populated with NaN values

Analysis of NaN values showed that the years 2003 – 2013 have very high (over 50% missing values), which caused this study to rely most heavily in the years 2014 2017.

US Zipcode database (link)
This database enabled the Wind Turbine (Windfarm) data file to be expanded to include the zip code, population density for each zip code, and median household income for each zip code. This was accomplished by using the windfarm's geocode/s) as the key to search the US Zipcode database.

Merged file:
The "expanded data file" (Windfarm + US Zipcode) was then joined to the Zillow Housing database based on zip code matching.

Explore the Data



As noted in the above image, the total zip codes dropped from 1,269 to only 20 that meet all criteria. The primary reason for this drop was that Zillow's housing data only had median sales price data for 3,797 zipcodes out of 42,000 zipcodes in the US.

Here is a summary of the steps that led to the final 20...

- Out of 1,269 zipcodes w/windfarms only 58 of these zip codes have Zillow housing data
- Of these 58 zip codes, only 46 zip codes are in states that have more than one zip code with a windfarm
- Of these 46 zip codes, there are only 20 zip codes twenty-five miles away without a windfarm that have Zillow housing data.

Even though only around 2% (20/1,269) of the windfarm's have corresponding housing data, there is sufficient data to continue with this study.

Continuing data exploration reveals the observations in the images below...

Exploring the Data (1 of 4)

- **States with highest and lowest count of zip codes with windfarms...**

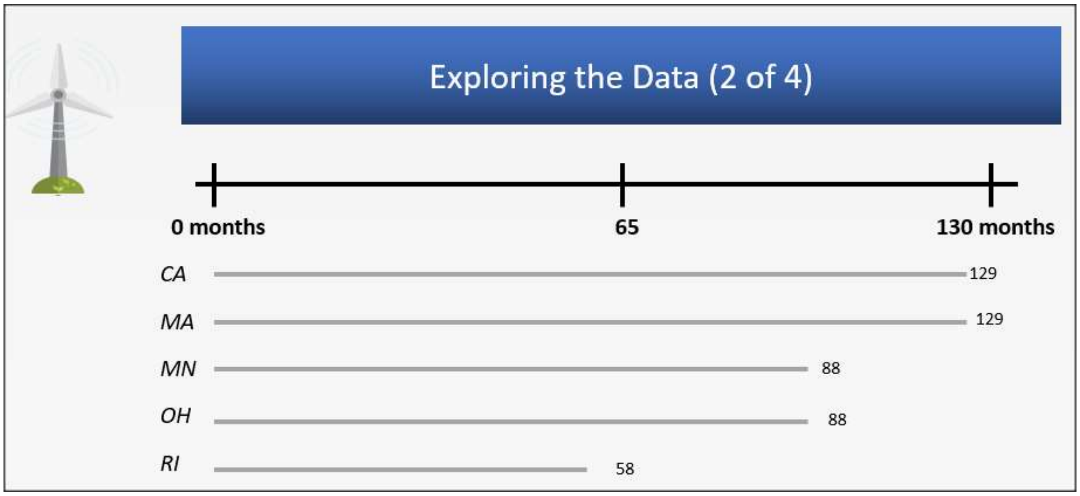
Highest 5

	state	zipcode_count
0	CA	19
4	MA	13
5	MN	6
7	OH	5
9	RI	3

Lowest 5

	state	zipcode_count
3	IL	1
6	NY	1
8	OK	1
11	UT	1
12	WA	1

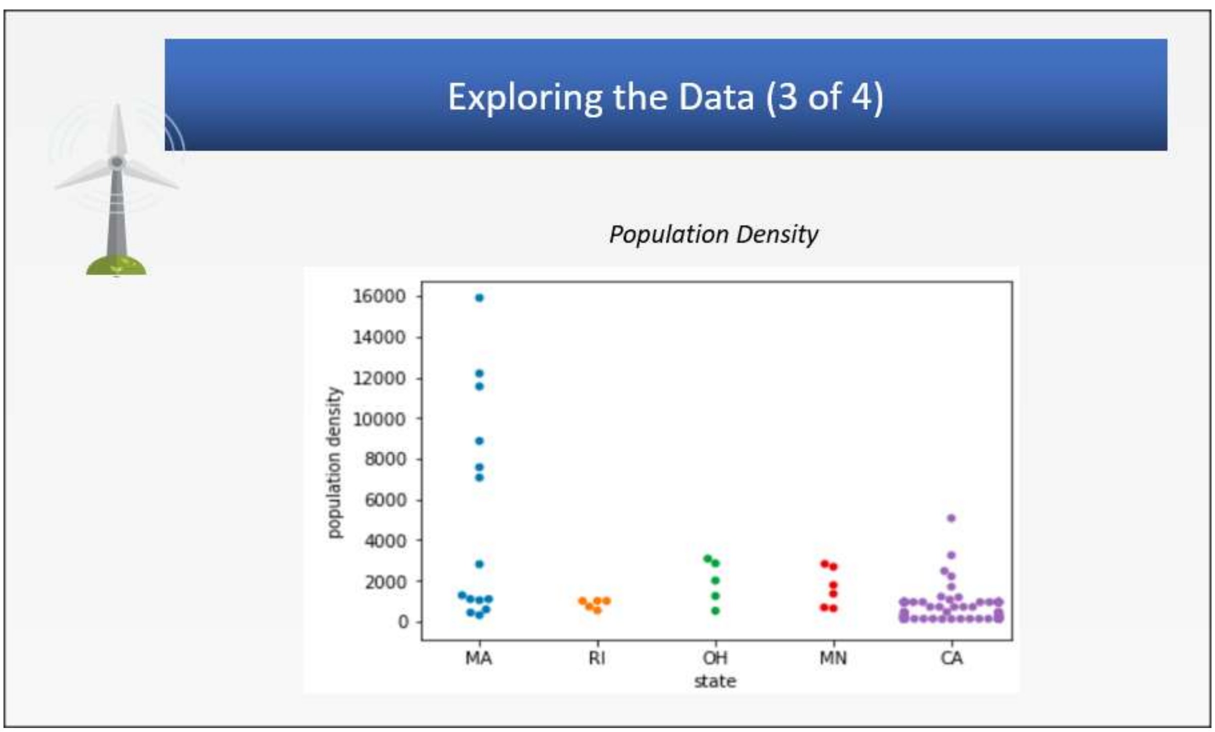
California tops the list for highest number of windfarms, while several states in which one would expect a higher count (i.e. Oklahoma and Illinois due to their flatter terrain and windy conditions) instead have a low count of only one windfarm. More likely though, these states only having one windfarm is a reflection of the previously noted poor amount of Zillow data available, rather than a true windfarm count of only one windfarm.



To Zillow's credit, when housing data is available it tends to be available for most of the months they track. An evaluation of housing data for the top 5 states in the image above reveals that out of a possible 130 housing values (months of Mar-2008 thru Aug-2018)...

- CA & MA have several windfarms w/129 months of housing data – very strong result
- MN & OH have several windfarms w/at least 88 months of housing data – sufficient result
- RI's windfarms have less than 58 months of data (less than 50%) – this isn't necessarily insufficient, but RI may not be a good state to keep in this study

Moving on to exploring population data and focusing on the same five states as the prior image reveals some interesting insights on population density...



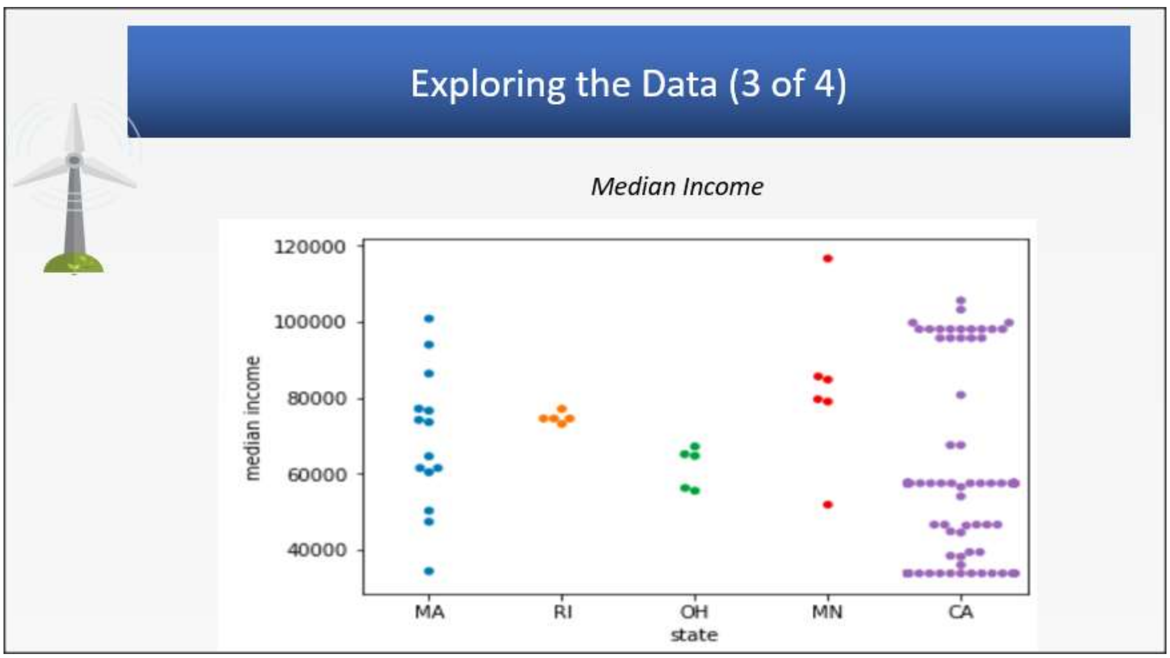
- Minimum density values range from 118 - 15,907 across windfarms with housing data available

- Visually inspecting the swarmplot illustrates...
  - that windfarms across the five states are generally located in areas of lower population density, with the exception of Massachusetts, which has a surprising number (6) of the windfarms located in moderately dense population areas.

  - the density minimum is very similar for Rhode Island, Ohio, and Minnesota, and the entire density range is very similar for Ohio and Minnesota.

  - Rhode Island has the tightest population density containing windmills and as noted above the density is a low score. The median income at around $80K, given the supposed rural location of the windmills, seems high.

Observations on median income are via visual inspection of the swarmplot below are...



- Minimum income with a windfarm is $33,682 and maximum income is $120,000

- a wide range of household incomes near windfarms for Massachusetts, Minnesota, and California; while for Rhode Island and Ohio the income range is much tighter.

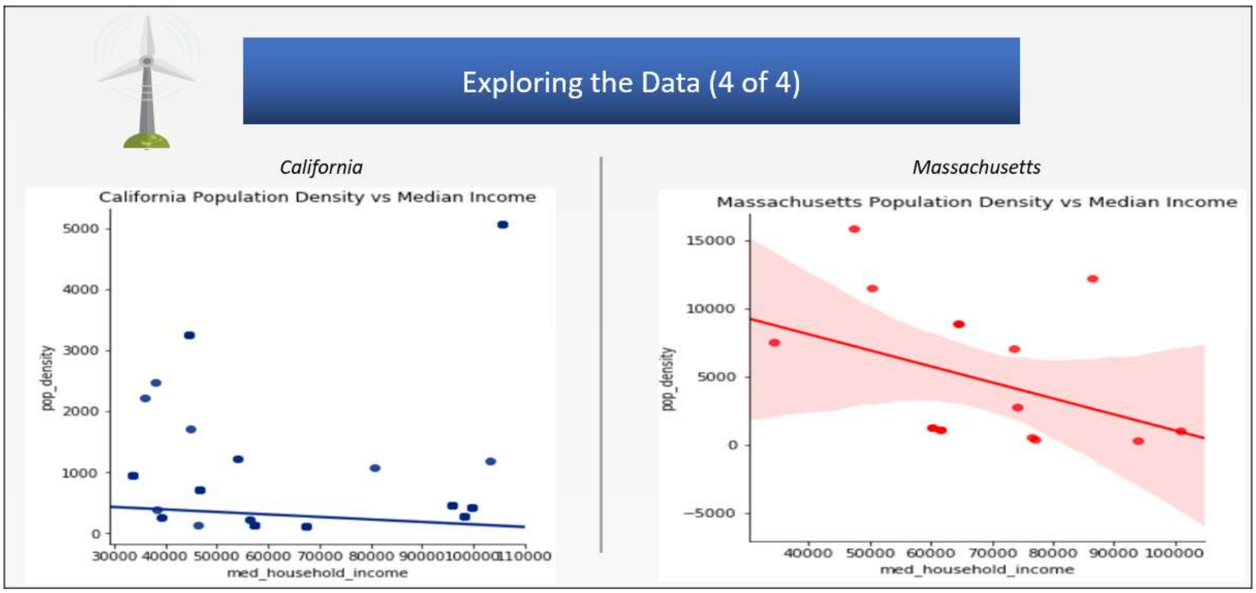
- four outlier datapoints of incomes greater than $100K (1 in Massachusetts, 1 in Minnesota, and 2 in California) that allude to windfarm/s located in higher income areas.

- Massachusetts has the most even distribution of windfarms located in lower to higher income areas.

- California has a majority of windfarms located in areas with income < $60,000, with a second smaller, but still significant number of windfarms located in areas with income > $90,000.

- As noted earlier Rhode Island and Ohio have the tightest population density range. These two states also have a tight income range. Minnesota's income range is also tight with the noted exception of one outlier data point. However, given the low count of windfarm's from these states included in this study (due to the constraint of available housing data) this may not be a valid observation.

The last exploration of the data is a comparison of MA and CA, the two states with the highest population densities. The plot below of population density vs median household income shows...



- MA income declines in rural areas, while CA has only a slight decline in income by population density.

- Comparing California to Massachusetts appears to illustrate that in California these two variables are not correlated

- In Massachusetts it appears that as income increases population density decreases. However, this result is questionable as the general trend in most states is that as population density increases (i.e. a metropolitan area) income increases due to a higher cost of living.

A comparison of population density vs median income was not done for Rhode Island, Minnesota, and Ohio because there is not sufficient data.

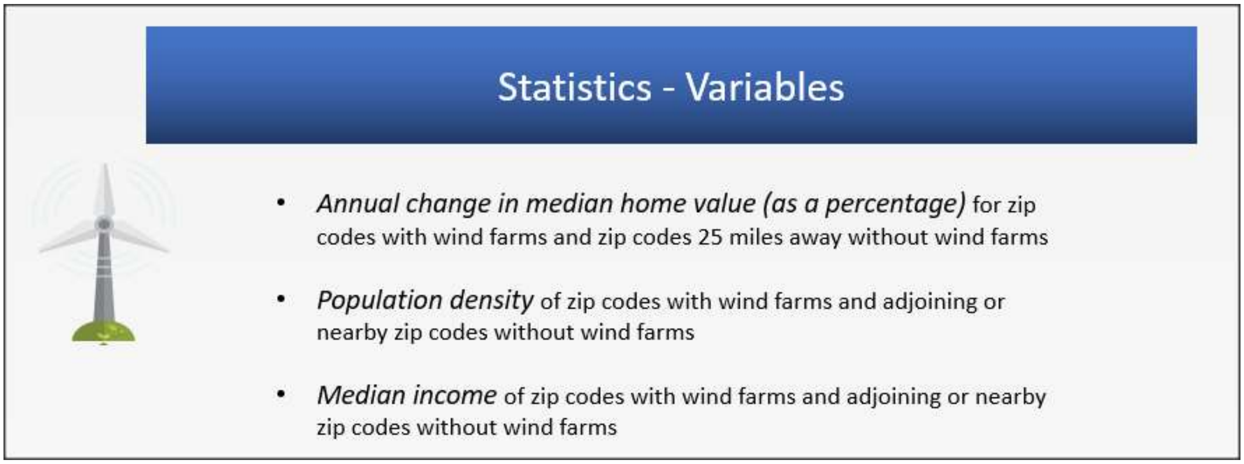
Statistics

Statistical inferences computed below are a mix of correlation coefficients, ECDF's, p-values, and confidence intervals. The objective being is a pre-machine learning analysis of what variables appear most significant to answer this project's question of...

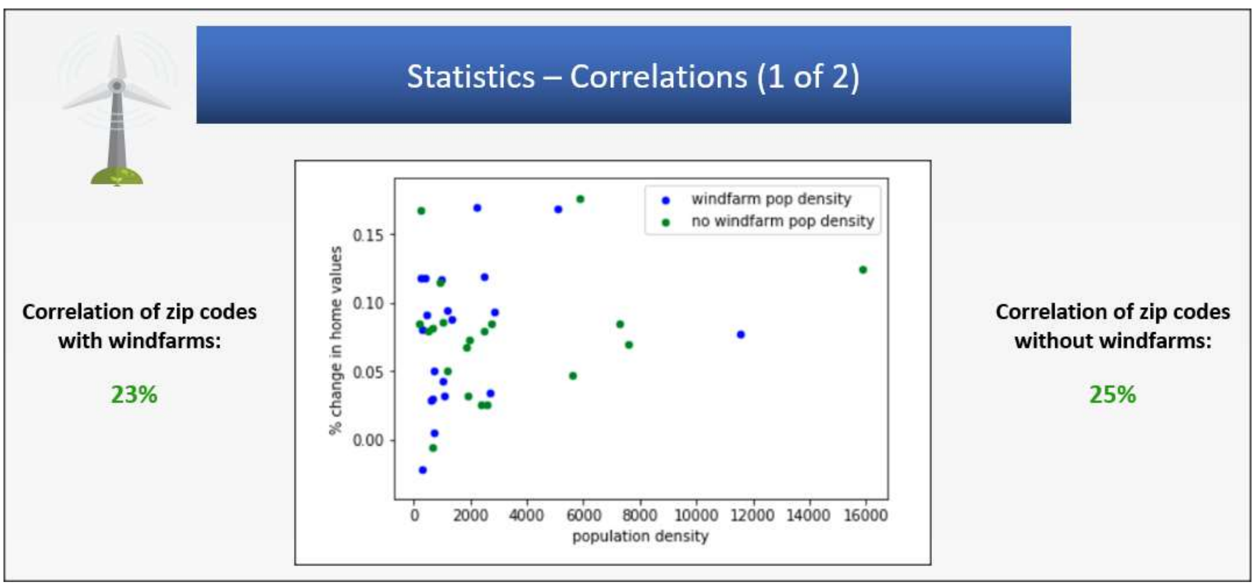
"Do windfarms impact the value of home prices?"

For this section the focus on the five states of CA, MA, RI, OH, and MN is continued. Several zipcodes from these states are extracted from the data via dataframes and used to calculate the inferences.

These are the variables currently being considered for the home value prediction model...

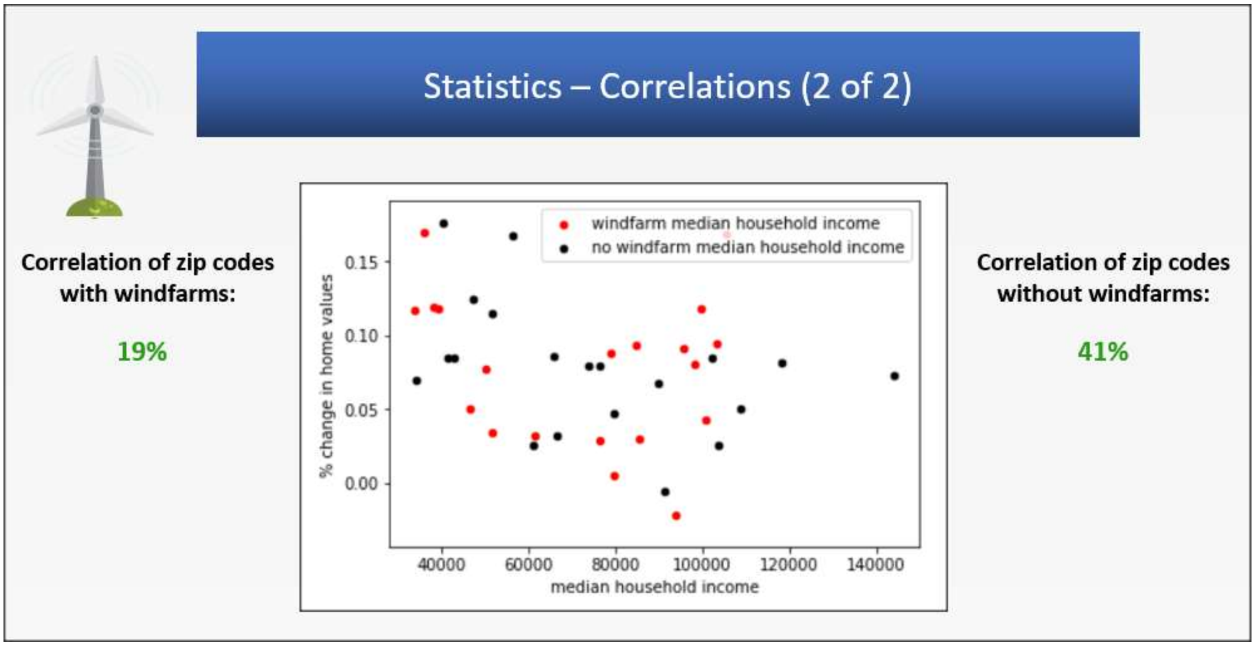


In the image below the correlation coefficient between Population Density & Mean % Change in Home Values differs by only 2%...



... which indicates that population density may not be a leading variable in forecasting the price of a home.

Moving on to the correlation coefficient between Median Income & Mean % Change in Home Values show a much higher difference of 22%...



... which indicates that median income may have a significantly higher prediction value than population density to predict the value of a home.

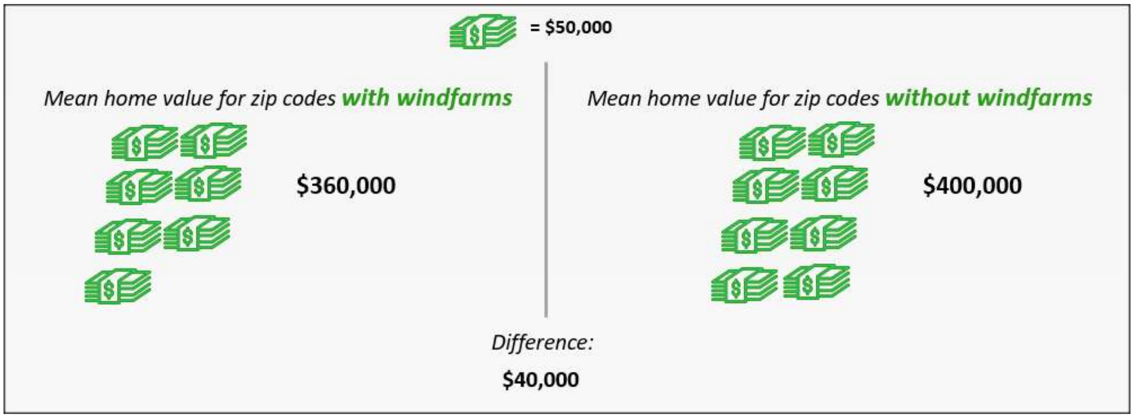
Using 2017 data, the image below shows the results of the hypothesis test...

#H0: mean home value for zipcodes with windfarms < mean home value for zipcodes w/out windfarms
#H1: mean home value for zipcodes with windfarms >= mean home value for zipcoes without windfarms

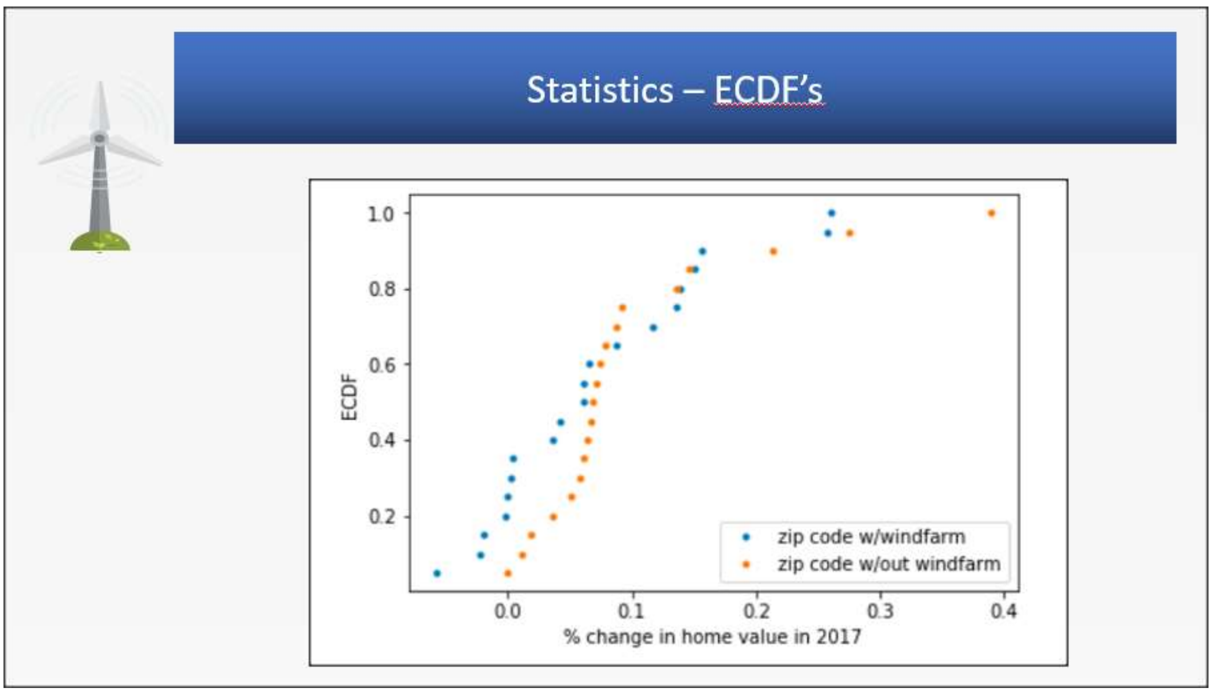
The p-value of 0.99 shows there is not a statistically significant difference in mean home values for zip codes with versus without windfarms.

The confidence interval of -$90K to $188K is a 95% confidence range for the mean difference in home values in 2017 in CA zipcodes with and without windfarms (see the image below)...

Capstone Project 1 Final Report
Predicting Wind Farms Impact on Home Values



The final statistical inference below continues the use of 2017 CA data to create ECDF's...



The results reveal that outliers may be causing the previously computed p-value and confidence interval to be misleading.

Machine Learning:

The algorithms below were employed to arrive at a model to predict home values...

1. Decision Tree
2. Random Forest
3. Ridge Regression
4. Lasso Regression

The data used by the machine learning algorithms was structured as follows...

1. Spans the years 2009-2017

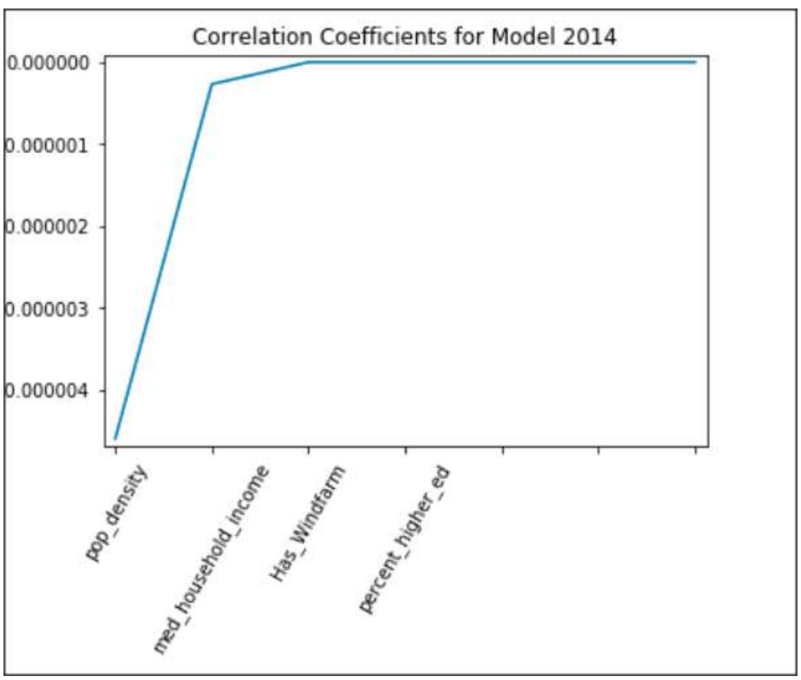
2. Breaks the zip codes into a series of datasets based on quality of housing data available. For example, a zip code with home values from 2009 – 2017 will be included in all datasets. A zip code with home value data starting in 2014 will be included in the 2014, 2015, etc... datasets, but not in the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

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...



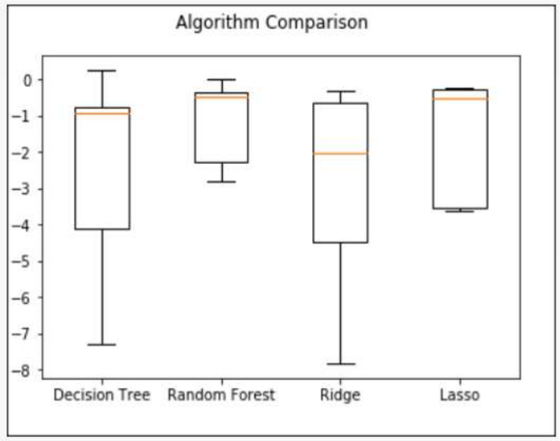
Each algorithm was analyzed individually and was also included in a comparison via the use of a box chart plot...

Machine Learning

- **Four algorithms were employed to research a model for predicting home values...**

Points of interest to learn from the box chart...

- All the algorithms produce a similar top, but Random Forest and Lasso appear have the highest top and a much tighter range than Ridge or Decision Tree. These appear to be the best models for further analysis.
- 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.



Summary:

In closing, the observations below summarize key conclusions from this study...

Summary

Restatement of Objectives...

1. Build a model that can be used to predict the change in value of a home
2. Assess if a windfarm in the vicinity of a home impacts the value of the home

Key Conclusions...

- Using a Random Forest Algorithm a model predicting home values with 77% accuracy was achieved
- The presence of a windfarm does not appear to have an impact on home values in the surrounding area
- The most significant predictor is population density, followed closely by median income.
- For 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 is windfarms may tend toward more rural areas and in rural areas median income may be less of a factor on home values.
- Not surprisingly, windfarms are primarily in rural areas, though several zip codes with high population density and high median incomes host windfarms