Capstone Project 1 Data Story
Predicting Wind Farms Impact on Home Values

Scope:

Location: There are 58 zip codes, distributed across 13 states, with windfarms that have corresponding housing data available.

Timeframe: Years 2008 – 2018. Note that limited housing data available prior to 2013 may result in this timeframe being modified to 2014 – 2018 as this study progresses.

The objective of this project is a regression algorithm that can be used to predict the change in value of a home based on proximity to windmills.

Explore the Data:

1. *Compare the original number of zip codes in the windfarm data to the number of windfarms that have corresponding information (a zip code match) with the Zillow housing data*
   original windfarm data count: 1,269
   merged windfarm data count: 58

   Even though only around 5% (58/1269) of the windfarm's have corresponding housing data, 58 zip codes across 13 states should be sufficient to continue with this study.

2. *Which states have the highest (and lowest) windmill counts*
   highest:

state	windfarm_count
CA	94
MA	14
MN	6
OH	5
RI	5

   lowest:

state	windfarm_count
IL	1
NY	1
OK	1
UT	1
WA	1

3. *Housing data availability: evaluate how many months of housing data are available for the above states with the highest and lowest mindmills count*
   An evaluation of housing data for the top 5 states revealed that out of a possible 130 housing values (months of Mar-2003 thru Aug-2018)...

   CA and MA have several windfarms with 129 months of housing data – very strong result
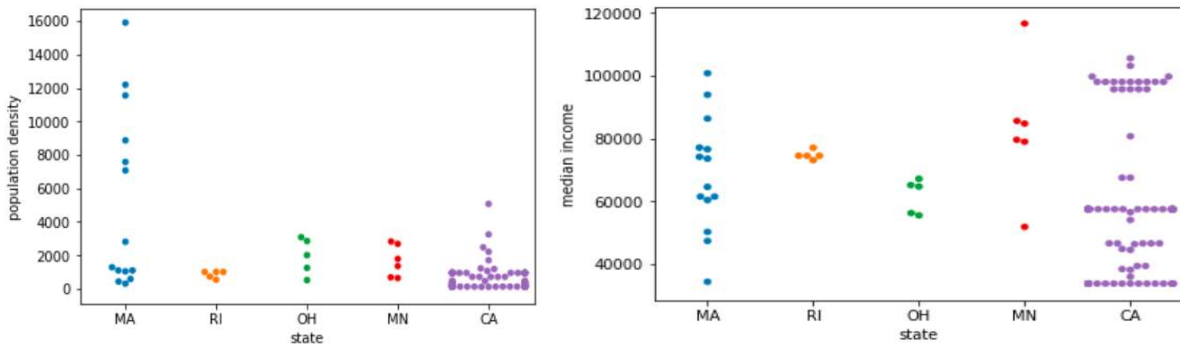
   MN and OH have several windfarms with at least 88 months of housing data – sufficient result

   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

4. *Swarmplot of population density and median household income for the five states*

For the population density,
Minimum density values range from 118 to 15,907 across all the windfarms with housing data availabe.

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.

For the median income,
Minimum income with a windfarm is $33,682 and maximum income is $120,000

Visually inspecting the swarmplot illustrates...
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.

5. *Compare two related quantities*

Capstone Project 1 Data Story
Predicting Wind Farms Impact on Home Values
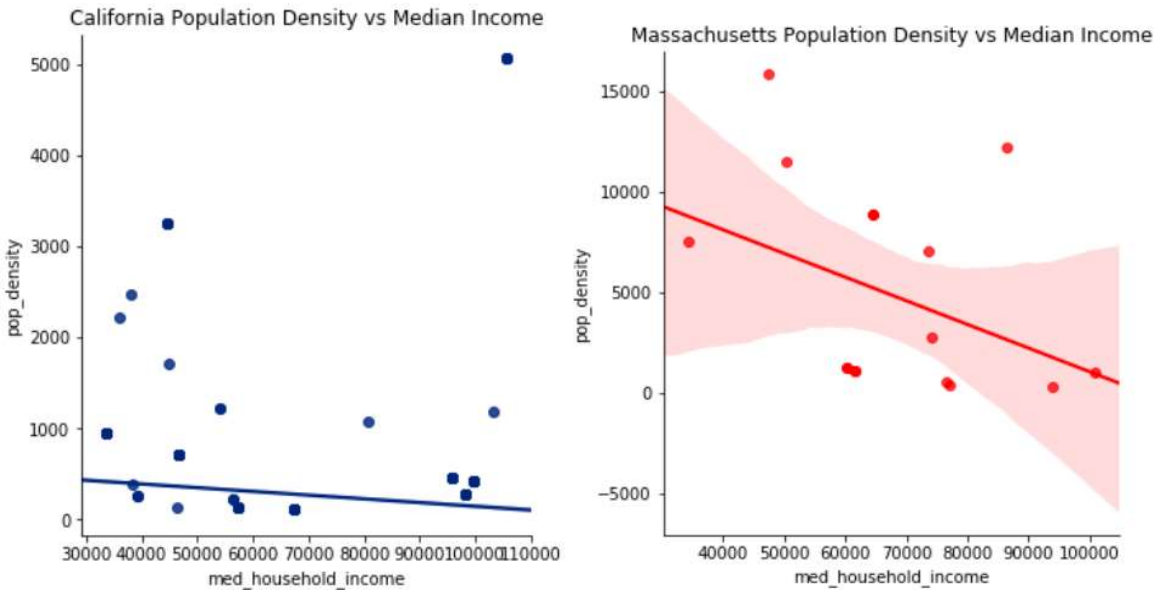
A comparison of population density vs median income...
was not done for Rhode Island, Minnesota, and Ohio because there is not sufficient data.

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.

Based on the exploration note any observations that will help build the regression model:

6. *Hypothesis to investigate further*
   Windfarms do not appear to be biased towards a lower or higher median income of households in the area surrounding the windfarm.

   Windfarms do not appear biased towards locations in higher or lower populated areas

7. *Questions this study won't explore*

Narrative…
8. Identify ways to engage the above information to tell a story about the impact of windmills on home values. The story will serve as key input to build a model that can predict future changes in home values