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

Objective:

Identify the data to clean (and how to clean) to complete the exploration described in the "Data Story".

Data Cleaning Steps:

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 came back as very clean…

Visual analysis shows that empty string fields in this file are populated with the word, "missing". However, an analysis of the data in the 10 columns needed comes 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 looks clean, but that is because missing data is populated with NaN values

Analysis of NaN values shows that the years 2003 – 2013 have very high (over 50% missing values). This may cause shift in analysis years from 2000 – 2010 to something more recent, i.e. 2014 – 2017

US Zipcode database (link)
Merged Wind Turbine database, Zillow Housing database, and US Zipcode database

The Wind Turbine (Windfarm) database was 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).

The "expanded data file" was then joined to the Zillow Housing database based on zip code matching.