Get area level data from -https://www.openintro.org/data/index.php?data=county\_complete.

State/Abbreviation data - https://github.com/jasonong/List-of-US-States/blob/master/states.csv

Will use this get the most important responses verbal (Wind capacity)/area

The data is from 5 sources

3 main – the proved by the prof

2 secondary – the links to where they can be found above

Wind Resources

Original each data point is a from a different testing facility/ turbine

Remove data from unknown state or county

Aggregate by state, county and take the mean of fraction\_of\_usable\_area, capacity, wind\_speed, capacity\_factor

Wind Turbine data

The where most of the data is from and it is originally at the turbine level

Eliminate all pre 2001 data and limit it to the continental USA

Make the variable for turbine capacity to MW

Removing the county from the county variable make merging possible.

When aggregating categorical data we take the mode

Wind ordinance data cleaning

This data is by county

Remove null data before 2001 and data that is not in the continental us

We also make a new data column that takes givens ordinances as a Boolean value

We get the take area of county data the prepare county for merage by making “XXXX county ” into “XXX” same a the last line of Wind Turbine data

Frist round of Merging

wind\_turbine\_main has t\_state which is the state Abbreviation while wind\_ordinance\_main, Area, and wind\_resource\_data has the name of State Spelled out.

We take the take wind\_ordinance\_main, Area, and wind\_resource\_data and merge it with state so that all data set now have state Abbreviation.

Then we merge wind\_turbine\_main and wind\_ordinance\_main by "Abbreviation" "County".

As ordinances only has data of place with ordinance we now make set any NULL values in the merged data set in that column as 0

Remove the fallowing variable from the merged data set that we do not use:

ordinance\_year

case\_id

faa\_ors

faa\_asn

usgs\_pr\_id

eia\_id

t\_fips

p\_name

retrofit

retrofit\_year

t\_img\_date

t\_img\_srce

xlong

ylat

Most are just ID values

Now we take the merged data set and make most of our final verbal by aggregating by county and state.

tot\_cap = sum(t\_cap)

avg\_p\_cap = mean(p\_cap)

avg\_hh = mean(t\_hh)

avg\_rd = mean(t\_rd)

ordinance = mean(ordinance)

## ^the aggregating function here does not matter as all enters per group have the same value

t\_manu = calculate\_mode(t\_manu)

turbines = n()

t\_model = calculate\_mode(t\_model)

t\_rsa = mean(t\_rsa)

t\_ttlh = mean(t\_ttlh)

avg\_t\_cap = mean(t\_cap)

This these are function used for aggregation in general for numeric it is mean, for categorical it is mode, and turbines in out of values aggregated

Lastly we take area of county and the information we took from wind resources and merge them with other data sets

Note:

We later remove avg\_t\_cap from the data set as would introduce a colinarty problem because turbines\*avg\_t\_cap = tot\_cap and tot\_cap is what we are trying to predict