Georgia Institute of Technology
Data Science Boot Camp, Team: Perfect Ten

04.13.19 | Project 01:
Pollution, Health and Income Relationships

Team Perfect Ten

Megan

Ivan

Lisa

Dan

Hypothesis

Team Perfect Ten's idea (created from limited evidence) is that there would be a correlation between size of state, amount of industrial plants within it, the immediately surrounding area's ability to absorb the pollution emitted and how this can negatively affect it's inhabitants level of health/mortality and income.

Motivation & Summary

CORE MESSAGE

While, there is an immediate relationship between population growth and the introduction of industrial plants to support it, we were compelled to identify the pollution they create and it's impact on health and income.

QUESTIONS ASKED

- Of the data sets available, what is the most recent year(s) we can utilize?
- Out of the 50 states, which ones have the most industrial plants?
- Out of the unique state facilities, how can we combine a matrix resulting in a weighted average of pollution emitted?
- Which state has the least land/water mass combined with the highest amount of pollution?
- With pollution state recognized, how can we include life expectancy and income datasets?
- Could facility type demonstrate a higher output of pollution and greater health/income impact?
- What about gender type, could there be an unequal impact on them?

QUESTIONS ANSWERED

Which is the state with worse amount of emissions?
We did a count of number of facilities. This gave us Texas but then begs the question because a high count of facilities is not a necessarily good indicator of how bad your toxic emissions are.
We weighed the count by including the ranking. We first got Virgin Islands, Guam, etc. but then we reverted the meaning of the ranking, from 1 being the worse to 1 being the lowest. It still begged the question, If you have a bigger state, then you will have more chances of having more emissions, spread out across more land than in a smaller state. Finally we normalized it by state size and getting ranking/square mile

LED TO MORE QUESTIONS/ FINDINGS

Based upon weighted average, Wisconsin was the state with the highest amount of toxic emissions?

- Which is the worse county in Wisconsin? Similar treatment than before.
- With a TRI/GHG rank of 18826.53, Milwaukee was found to be the county with by far the most toxic emissions due to facilities within the state of Wisconsin. The county of Waukesha came in 2nd with the most toxic emission with a ranking of 3296.94.

What is the county with the greatest life expectancy?
What's the county with the greatest median income?
Is there any correlation between Life expectancy vs toxic emissions?
Is there any correlation between Median income vs toxic emissions?
Is there any correlation between Life expectancy vs Median income?
What other interesting facts we have unraveled?
100% not reported only in 2 counties.
The non-reported facilities are concerning because it could mean anything?
(Clark vs Washburn example).

Does the low reporting patterns of industrial plants pollution in areas of high mortality rates mean they have something to hide?

PERFECT TEN

Questions & Data

Online data sets and state size my provide a better picture of the housing preferences driven by price and link to areas of income

Gender, household price, geo location, frequency of industrial plants, and regional health status may tell us more about the profile of our residents

Information collected about the concentrated areas of pollution can help us come up with context that may influence health and mortality rates in US.

Regional location of concentrated pollution, or lack thereof, may indicate certain preference and relevance of housing location choice

We can munge, and visualize the data and attempt to establish meaning behind areas of the United States that impact health, pollution, income and even a type of industrial plant to avoid.

The choice in location to live in respect to the industrial plant generators of pollution can tell us more about the appropriate, potential content and format to engage, inform, and empower residents to perhaps make a better housing choice for them and family

PERFECT TEN

Elements

Target Population

Review of population in reference to gender, frequency of industrial plants, amount of pollution generated, mortality rate, and household income

Data Samples

Collection of data with accuracy and reputation top of mind

Desired Criteria

Input of criteria based upon example of smallest state with highest frequency of industrial plants, impact on mortality rate and household income

Munging/Learning

Munging driven by outcomes, outliers, and key data points that illuminate meaningful insights

PERFECT
P X
TEN

Project Proposal

DATA TYPE

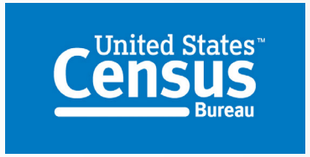
Integrity, accuracy and reputation of data source was prioritized in line with the eligible timeline that we could pull through a variety of datasets:

- Institute for Health Metrics and Evaluation
- EPA
- Plotly
- Golden Oaks Research Group
- 2010 Census

THE KINDS OF QUESTIONS YOU'LL BE ASKING OF THAT DATA

- How/why can smaller states have such a high number of industrial plants resulting–can land/water sustainably absorb this?
- Is there a direct connection to densely polluted areas of US and increase in mortality rates?
- How would all of this impact a woman versus a man's health?
- How does this impact the housing prices in proximity to the highest areas of pollution?

Reliable Data Sources



Data Explore & Clean-up

[51…
```
df3 = df2.groupby('State').count()
df4 = df2.groupby('State').sum()
```

[52…
```
len(df3.index)
```

[52… 55

[53…
```
weight_rank = []

for i in range(0,len(df3)):
    item1 = df3.iloc[i,0]
    item2 = df4.iloc[i,2]
    weight_rank.append(item1*item2/1000)

data = {'State': df3.index, 'Weighted Rank': weight_rank}
```

[54…
```
len(weight_rank)
```

[54… 55

[55…
```
df5 = pd.DataFrame(data)
```

[56…
```
df5.sort_values(by = 'Weighted Rank', ascending=False).head()
```

[56…

	State	Weighted Rank
46	TX	59593.679200
37	OH	40873.741243
15	IL	25467.078530
40	PA	24929.193643
4	CA	22426.968666

[57…
```
states_size = pd.read_csv('states size.csv')
states_size
```

[57…

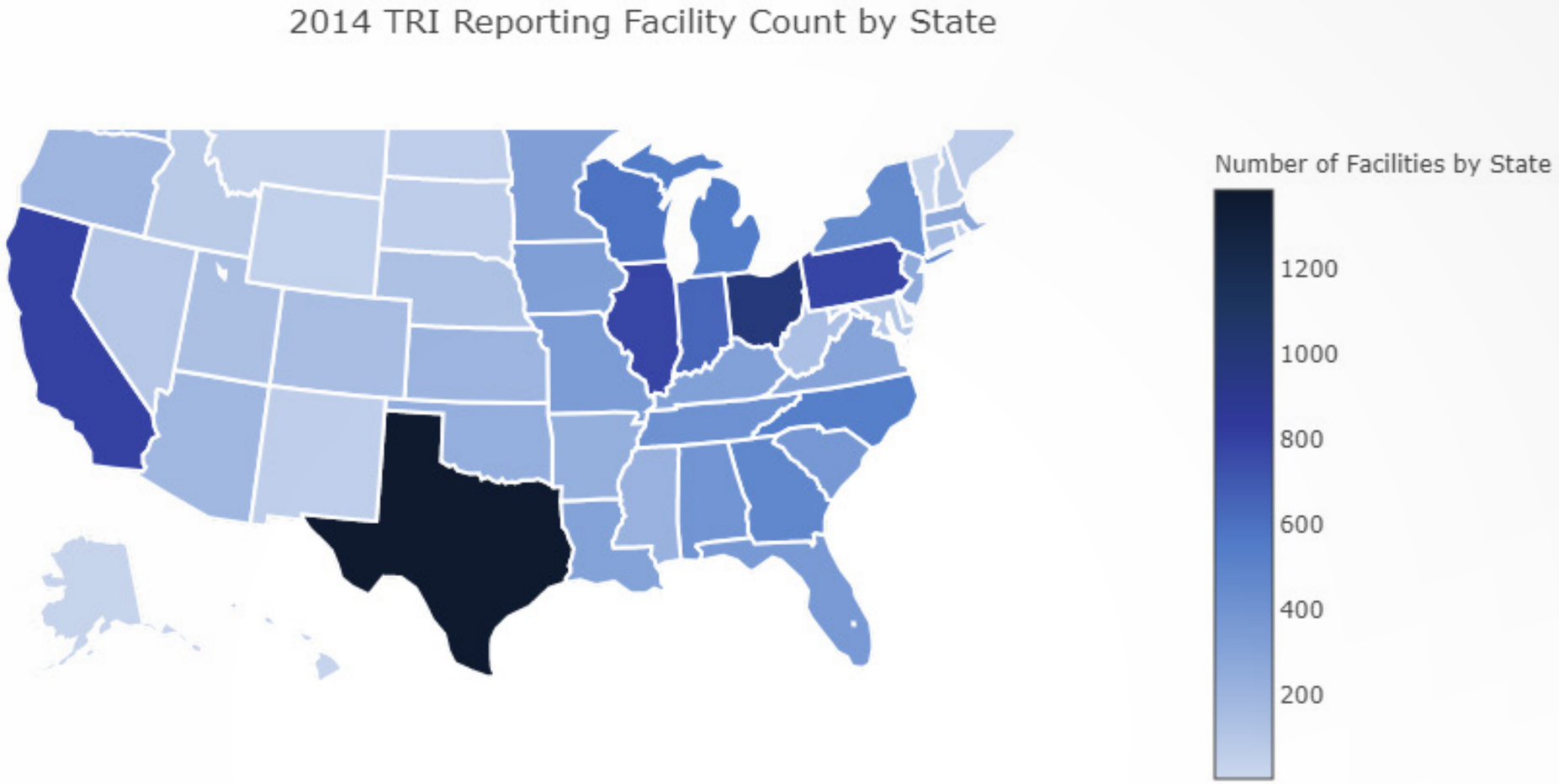
	State	Rank	sq mi	km^2	Rank.1	sq mi.1	km^2.1	% land	Rank.2	sq mi.2	km^2.2	% water
0	Alaska	1.0	665,384.04	1,723,337	1.0	570,640.95	1,477,953	85.76%	1.0	94,743.10	245,384	14.24%
1	Texas	2.0	268,596.46	695,662	2.0	261,231.71	676,587	97.26%	8.0	7,364.75	19,075	2.74%
2	California	3.0	163,696.32	423,972	3.0	155,779.22	403,466	95.16%	6.0	7,915.52	20,501	4.84%
3	Montana	4.0	147,039.71	380,831	4.0	145,545.80	376,962	98.98%	26.0	1,493.91	3,869	1.02%
4	New Mexico	5.0	121,590.30	314,917	5.0	121,298.15	314,161	99.76%	49.0	292.15	757	0.24%
5	Arizona	6.0	113,990.30	295,234	6.0	113,594.08	294,207	99.65%	48.0	396.22	1,026	0.35%
6	Nevada	7.0	110,571.82	286,380	7.0	109,781.18	284,332	99.28%	36.0	790.65	2,048	0.72%
7	Colorado	8.0	104,093.67	269,601	8.0	103,641.89	268,431	99.57%	44.0	451.78	1,170	0.43%
8	Oregon	9.0	98,378.54	254,799	10.0	95,988.01	248,608	97.57%	20.0	2,390.53	6,191	2.43%
9	Wyoming	10.0	97,813.01	253,335	9.0	97,093.14	251,470	99.26%	37.0	719.87	1,864	0.74%
10	Michigan	11.0	96,713.51	250,487	22.0	56,538.90	146,435	58.46%	2.0	40,174.61	104,052	41.54%
11	Minnesota	12.0	86,935.83	225,163	14.0	79,626.74	206,232	91.59%	9.0	7,309.09	18,930	8.41%
12	Utah	13.0	84,896.88	219,882	12.0	82,169.62	212,818	96.79%	17.0	2,727.26	7,064	3.21%

[70…
```
df5.sort_values(by = 'Final Ranking', ascending=False)
```

[70…

	State	Weighted Rank	Size sq mi	Final Ranking
52	WI	15171.927629	1975.57	7679.772232
46	TX	59593.679200	8722.58	6832.116094
48	VA	3631.891323	2488.72	1459.341076
53	WV	843.294770	732.93	1150.580233
37	OH	40873.741243	36419.55	1122.302204
40	PA	24929.193643	24230.04	1028.854828
45	TN	6451.988891	9349.16	690.114287
43	SC	5111.896045	10554.39	484.338370
51	WA	2478.754885	5324.84	465.507862
15	IL	25467.078530	77347.81	329.254035
50	VT	17.603111	68.34	257.581370
16	IN	17177.080608	77115.68	222.744332
36	NY	8956.288520	40407.80	221.647517
23	MI	12984.518682	59425.15	218.502077
29	NC	10327.105297	52420.07	197.006706
4	CA	22426.968666	121590.30	184.447021
54	WY	90.447904	581.05	155.662858
47	UT	806.216058	5543.41	145.436844
24	MN	5028.520818	57913.55	86.828053
25	MO	4868.992129	56272.81	86.524773
10	GA	7943.856416	96713.51	82.138022
33	NJ	2756.112358	44825.58	61.485258
38	OK	2127.020451	35379.74	60.119731
18	KY	4194.167677	70698.32	59.324856
13	IA	4836.556630	83568.95	57.875044
39	OR	1792.885983	32020.49	55.991835
19	LA	3148.772001	69898.87	45.047538
20	MA	3014.211552	69706.99	43.241166
9	FL	3990.049537	97813.01	40.792626
27	MS	1700.404215	53819.16	31.594774
17	KS	1689.730258	71297.95	23.699563
1	AL	5497.128737	268596.46	20.466125
44	SD	192.347880	9616.36	20.002151
31	NE	784.481755	48431.78	16.197665
42	RI	145.375683	10931.72	13.298519

Data Explore & Clean-up



Data Analysis Process

```
[73...]
work_df = df.loc[df['State'] == 'WI']
work_df.sort_values(by = "Rank_TRI_14").head()
```

[73...]

	Unique ID	FacilityName	Rank_TRI_14	Rank_GHG_14	Latitude	Longitude	LocationAddress	City	State	ZIP	County	FIPScode	PrimaryNAICS	SecondPrimaryNAICS	ThirdPrimaryN
25921	110017413547	WISCONSIN RAPIDS PULP MILL	96.0	1126.0	44.40400	-89.825500	950 4TH AVE N	WISCONSIN RAPIDS	WI	54495	WOOD	55141.0	322110	322121.0	
25556	110000420973	THILMANY PAPER MILL	151.0	1232.0	44.28290	-88.251800	600 THILMANY RD	KAUKAUNA	WI	54130	OUTAGAMIE	55087.0	322121	NaN	
25022	110013863275	GEORGIA-PACIFIC CONSUMER PRODUCTS LP	201.0	694.0	44.49250	-88.032300	1919 S BROADWAY	GREEN BAY	WI	54304	BROWN	55009.0	322121	NaN	
25922	110000573692	BIRON MILL	284.0	1124.0	44.42890	-89.781700	621 BIRON DR	WISCONSIN RAPIDS	WI	54494	WOOD	55141.0	322121	NaN	
25923	110000544233	WATER QUALITY CENTER	361.0	3654.0	44.42328	-89.831663	2811 5TH AVE N	WISCONSIN RAPIDS	WI	54495	WOOD	55141.0	322121	NaN	

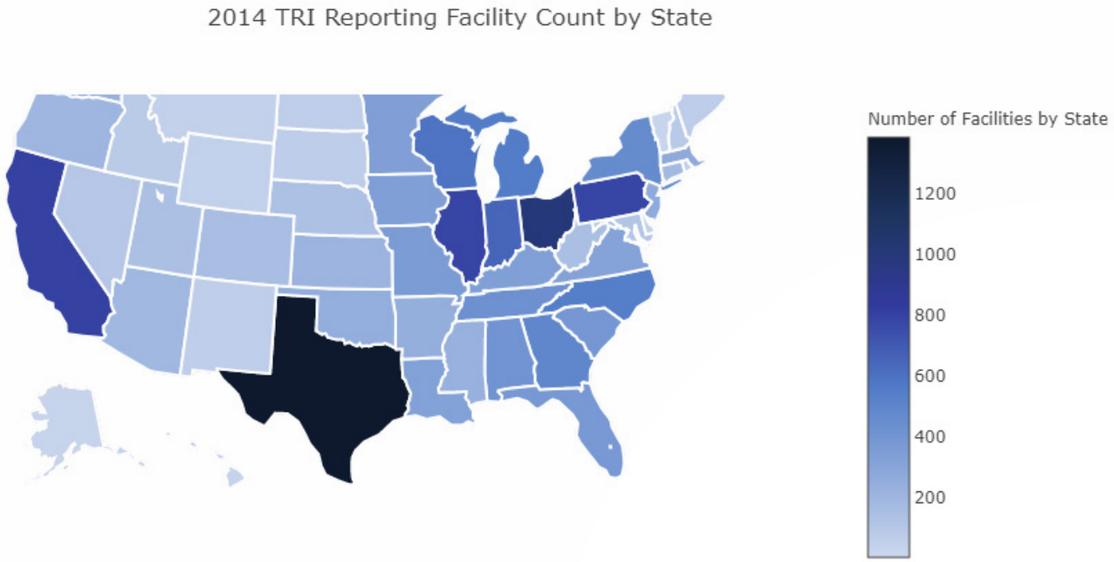
```
[72...]
work_df.groupby('IndustryType').count()
# work_df.groupby('County').count()
```

[72...]

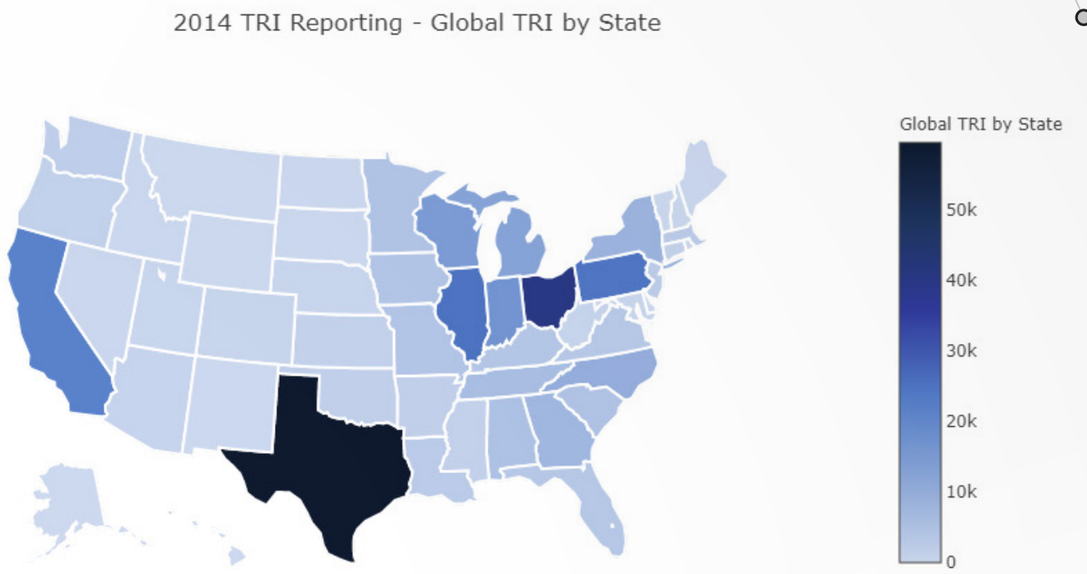
IndustryType	Unique ID	FacilityName	Rank_TRI_14	Rank_GHG_14	Latitude	Longitude	LocationAddress	City	State	ZIP	County	FIPScode	PrimaryNAICS	SecondPrimaryNAICS	ThirdPrimaryNAICS	Parent_
Chemicals	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	
Metals	6	6	6	6	6	6	6	6	6	6	6	6	6	2	0	
Minerals	8	8	7	8	8	8	8	8	8	8	8	8	8	1	0	
Minerals,Waste	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	
Other	17	17	10	17	17	17	17	17	17	17	17	17	17	3	0	
Other,Suppliers of CO2	3	3	3	3	3	3	3	3	3	3	3	3	3	0	0	
Other,Suppliers of CO2,Waste	2	2	2	2	2	2	2	2	2	2	2	2	2	0	0	
Other,Waste	3	3	3	3	3	3	3	3	3	3	3	3	3	1	0	
Petroleum Product Suppliers,Refineries	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	
Petroleum and Natural Gas Systems	1	1	0	1	1	1	1	1	1	1	1	1	1	0	0	
Power Plants	40	40	17	40	40	40	40	40	40	40	40	40	40	2	0	
Pulp and Paper	20	20	15	20	20	20	20	20	20	20	20	20	20	1	0	
Pulp and Paper,Suppliers of CO2,Waste	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	
Pulp and Paper,Waste	8	8	7	8	8	8	8	8	8	8	8	8	8	0	0	
Waste	29	29	0	29	29	29	28	29	29	29	29	29	29	0	0	

Nationwide Count vs. Weighted Average

COUNT

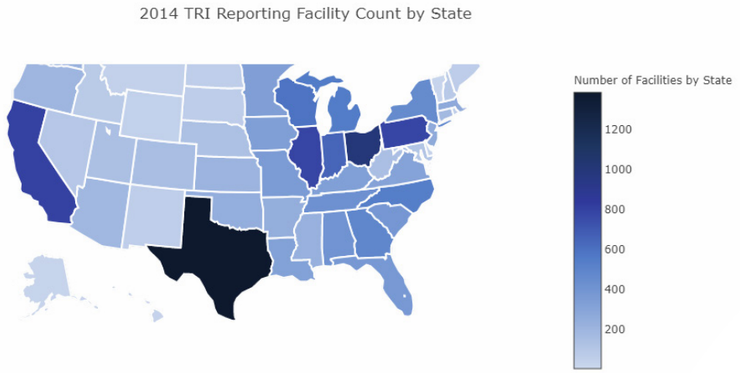


WEIGHT AVE.

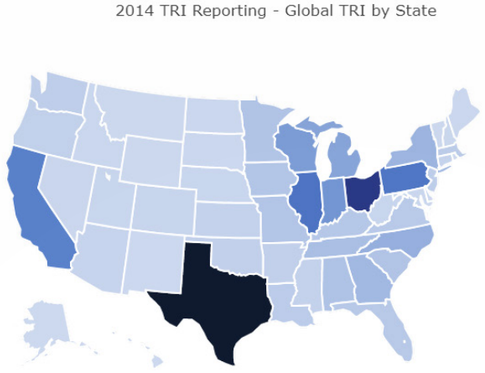


Nationwide State Comparison

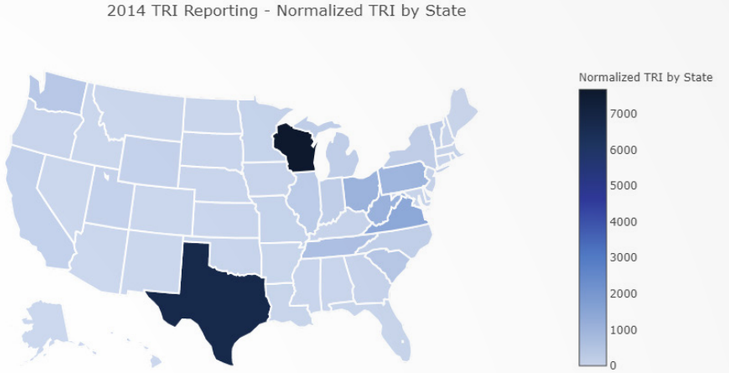
COUNT



WEIGHT AVE.



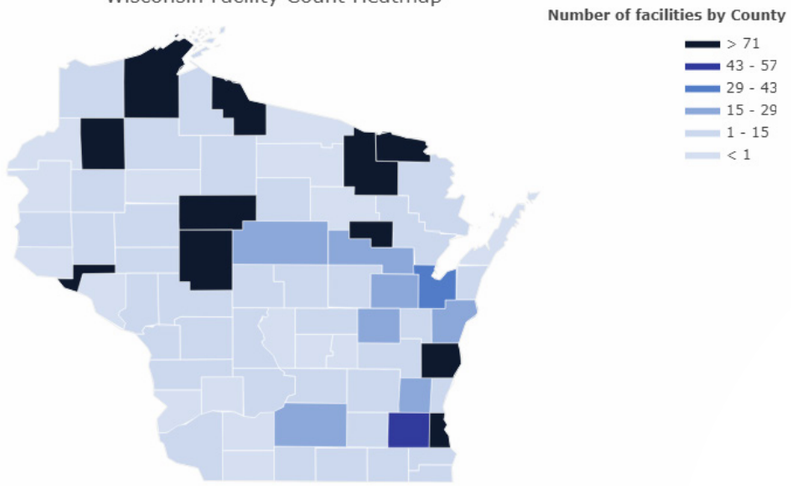
NORMALIZED WEIGHT AVE.



Wisconsin Comparison

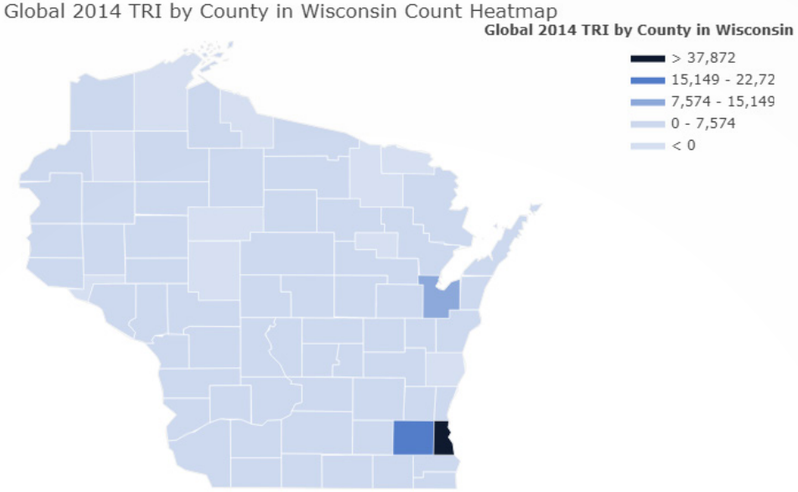
COUNT

Wisconsin Facility Count Heatmap



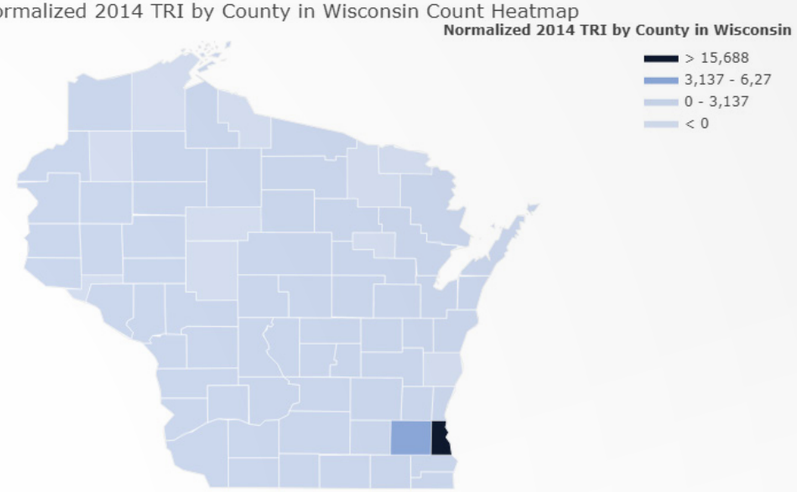
WEIGHT AVE.

Global 2014 TRI by County in Wisconsin Count Heatmap



NORMALIZED WEIGHT AVE.

Normalized 2014 TRI by County in Wisconsin Count Heatmap



Conclusions

FINDING

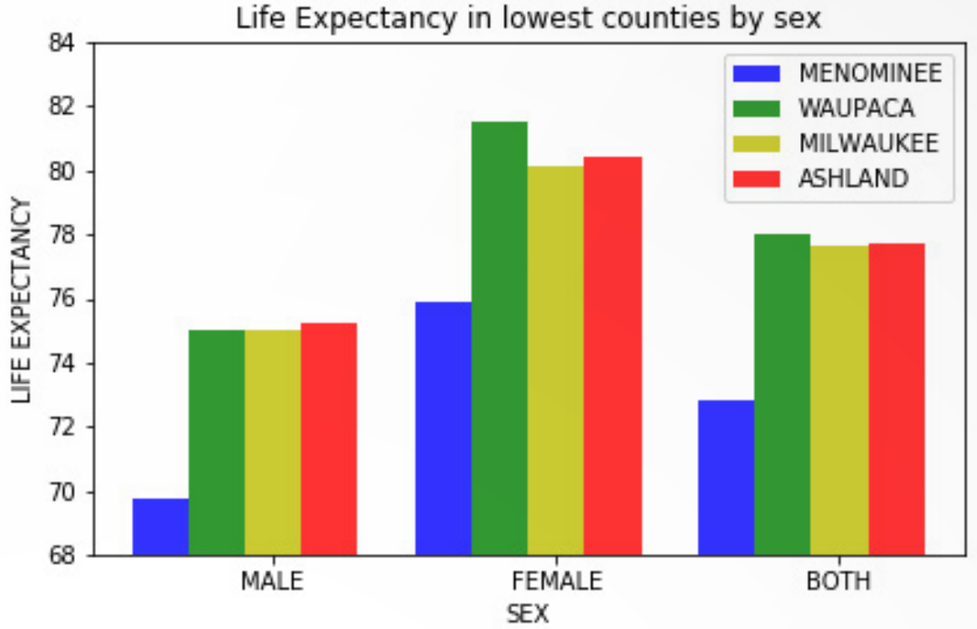
Women outliving men in all categories, (pollution, income, county)

QUESTION ASKED OF DATA

Wanted to see where the lowest life expectancies are located in Wisconsin by county?

FOUND/DISCOVERED

- Women live longer than men in Menominee County by 9.35% (comparatively significant)
- Women live longer than men in Waupaca County by 9.3%
- Women live longer than men in Wisconsin by 6.49%



*Menominee was the lowest rank in the weighted average

```
[15…
#drops years that are not 2014
healthdf.drop(healthdf.index[healthdf['year_id'] != 2014], inplace = True)
#drop mortality risk
healthdf.drop(healthdf.index[healthdf['measure_name'] != 'Life expectancy'], inplace = True)

#formatting
healthdf['location_name'] = healthdf['location_name'].str.upper()
healthdf['location_name'] = healthdf['location_name'].str.replace('COUNTY','')

#Pivot Table of Data Frame
healthpivot = pd.pivot_table(healthdf,index=['FIPS','location_name','sex'],
                             values=['val'],
                             aggfunc={'val':np.mean})
#show lowest life
healthpivot = healthpivot.nsmallest(5,'val')

healthpivot
```

[15…

			val
FIPS	location_name	sex	
55078	MENOMINEE	Male	69.800259
		Both	72.814850
55135	WAUPACA	Male	75.007037
55079	MILWAUKEE	Male	75.048078
55003	ASHLAND	Male	75.221359

```
[15…
#seperate data frames
healthboth_df = healthdf.loc[healthdf['sex'] == 'Both']
healthma_df = healthdf.loc[healthdf['sex'] == 'Male']
healthfe_df = healthdf.loc[healthdf['sex'] == 'Female']
healthfe_df
```

[15…

	measure_id	measure_name	location_id	location_name	FIPS	sex_id	sex	age_id	age_name	year_id	metric	val	upper	lower
69	26	Life expectancy	572	WISCONSIN	55	2	Female	161	0	2014	Years	81.898503	81.982344	81.810366
174	26	Life expectancy	3670	ADAMS	55001	2	Female	161	0	2014	Years	81.036637	81.817024	80.282969
279	26	Life expectancy	3622	ASHLAND	55003	2	Female	161	0	2014	Years	80.386521	81.215060	79.524497
384	26	Life expectancy	3629	BARRON	55005	2	Female	161	0	2014	Years	81.817434	82.401250	81.228511
489	26	Life expectancy	3620	BAYFIELD	55007	2	Female	161	0	2014	Years	81.991371	82.796236	81.171766
594	26	Life expectancy	3667	BROWN	55009	2	Female	161	0	2014	Years	82.474285	82.867707	82.076752

Conclusions

FINDING

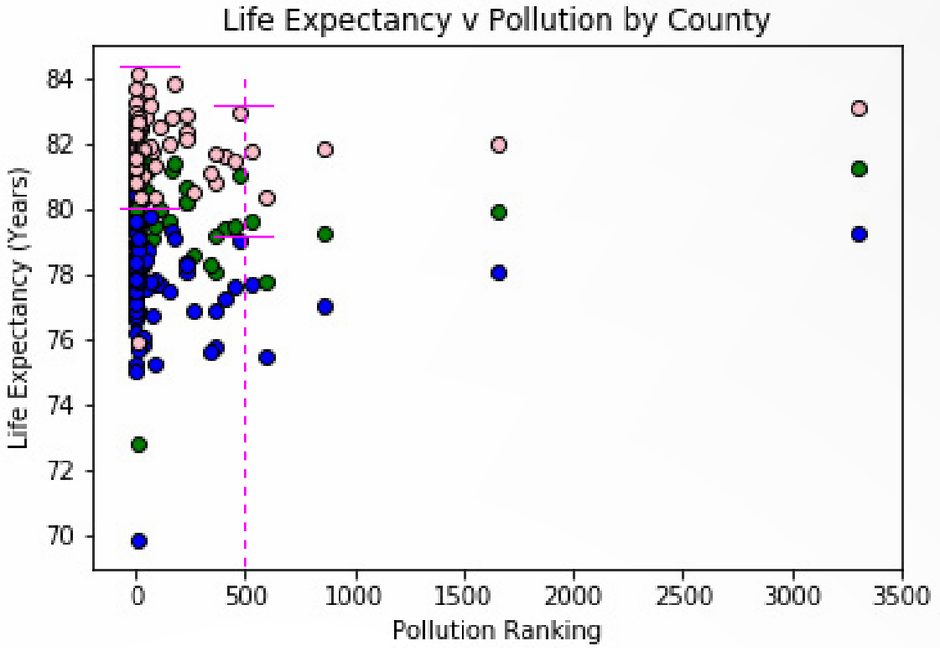
We expected to plot pollution's impact on life expectancy by gender

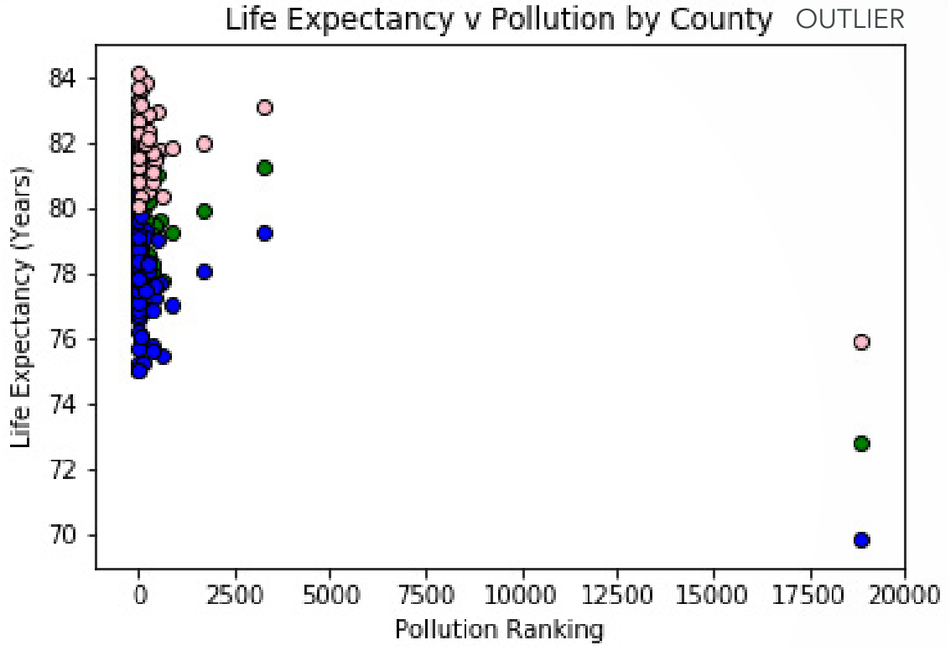
QUESTION asked of data

Find the correlation between life expectancy and pollution.

FOUND/REINFORCED

Women, with respect to pollution live longer than men in the state of Wisconsin





Conclusions

FINDING

We expected to see a positive correlation between income and life expectancy

QUESTION asked of data

Is there a trend between how much you make and how long you live in Wisconsin?

FOUND/DISCOVERED

Within a 67.3% difference in income, women still outlive men by a minimum of 7.5 years.

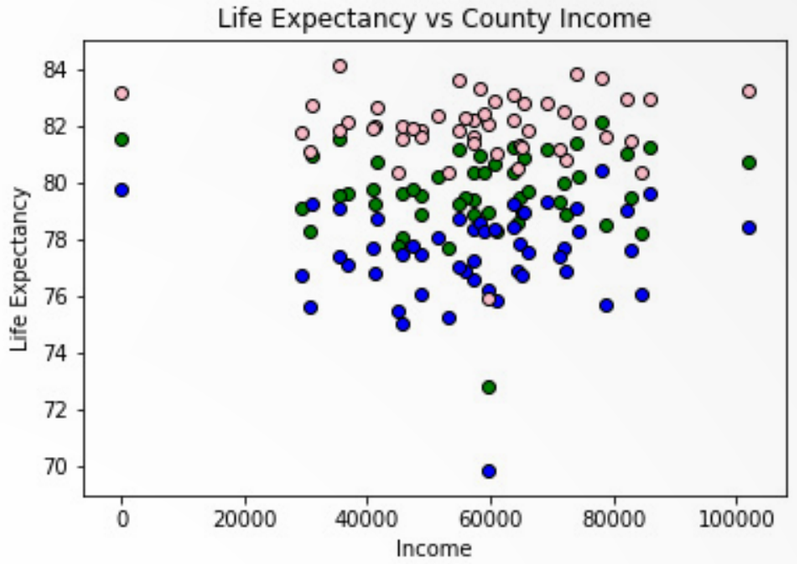
We pulled out life expectancy versus country to prove pollution is the driving factor in life expectancy, (not divorce).

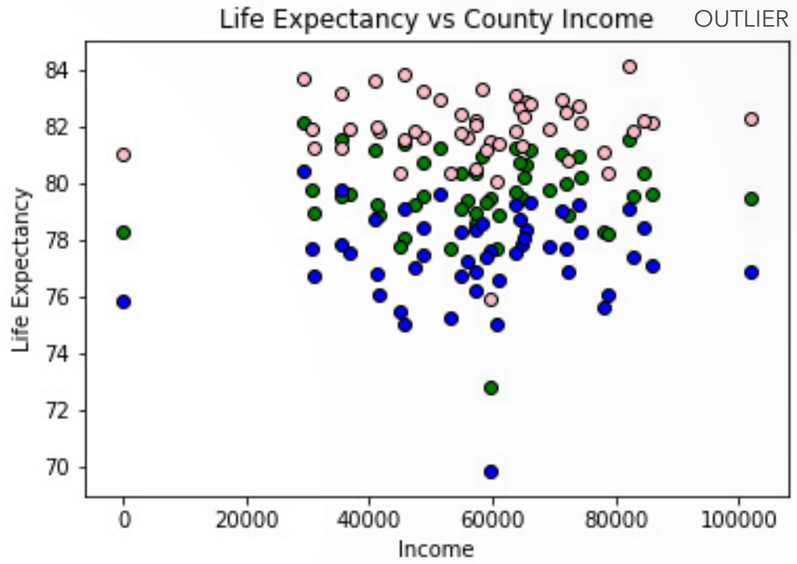
What we found is that income is not a driver of life expectancy in Wisconsin.

[16…

```
#3) Scatter plot: Life Expectancy vs County names (ordered by income)

# Create a title, x label, and y label for our chart
plt.title("Life Expectancy vs County Income")
plt.ylabel("Life Expectancy")
plt.xlabel("Income")


y_axis4 = healthboth_df['val'].sample(n=50, random_state=1)
y_axis5 = healthma_df['val'].sample(n=50, random_state=1)
y_axis6 = healthfe_df['val'].sample(n=50, random_state=1)
x_axis2 = incomepivot['Mean']

plt.scatter(x_axis2, y_axis4, marker="o", facecolors="green", edgecolors="black")
plt.scatter(x_axis2, y_axis5, marker="o", facecolors="blue", edgecolors="black")
plt.scatter(x_axis2, y_axis6, marker="o", facecolors="pink", edgecolors="black")

plt.savefig("Images/LifevCountyOutlier.png")


print(len(healthdf['val']))
print(len(incomepivot['Mean']))
```

```
219
50
```





Conclusions

FINDING

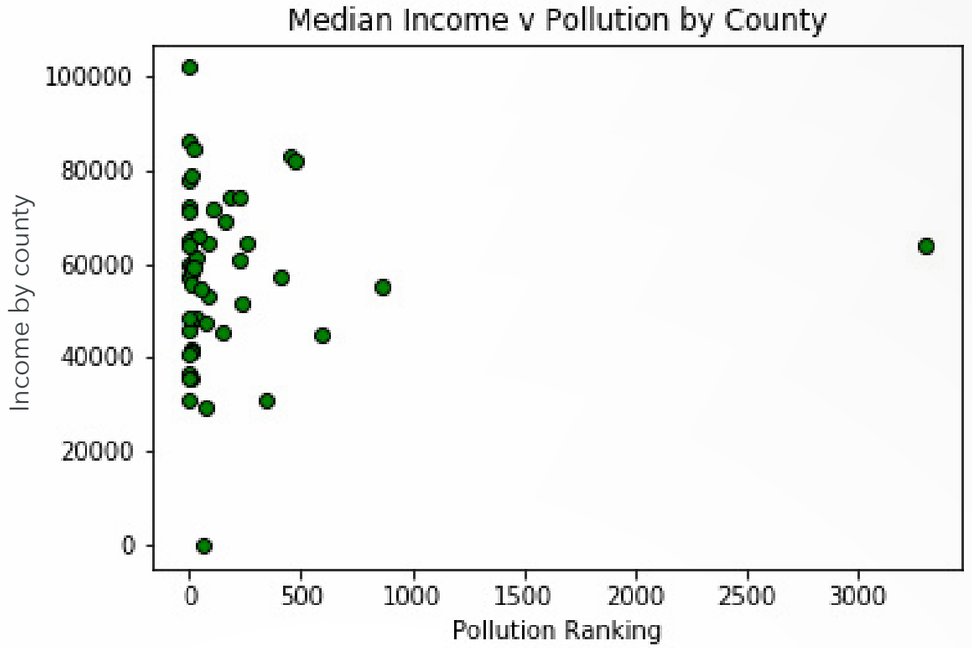
Uncover some evidence that utility plants are located within lower income areas out of convenience and cost versus a consideration for residents

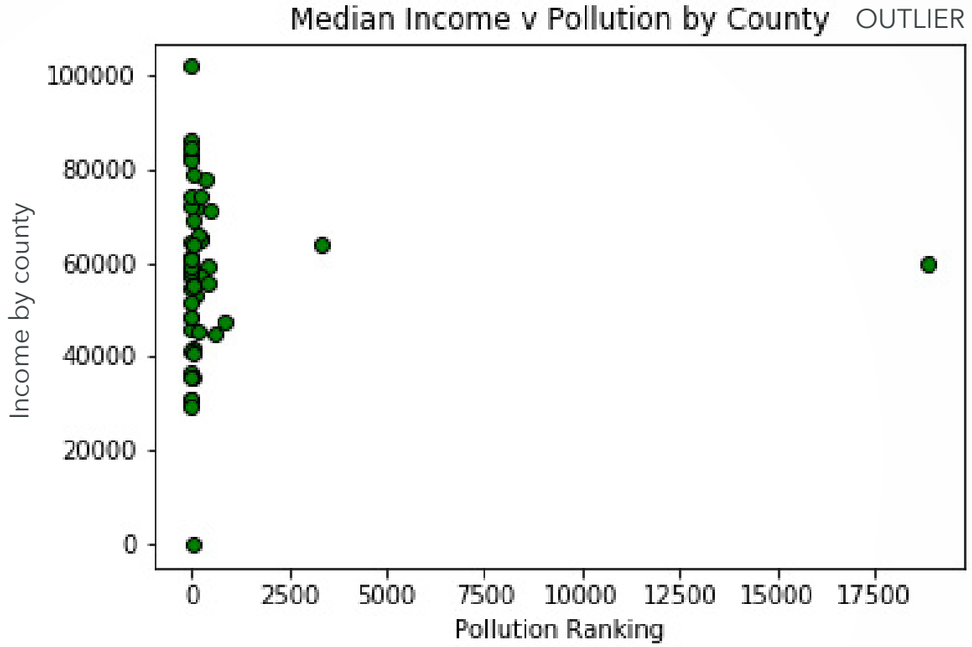
QUESTION asked of data

Does the location of utility plant installation is driven by the primary criteria of lower land values and henceforth a lower level of property taxes collected.

FOUND/DISCOVERED

To some extent there is a relationship, but would have to be more explored in smaller regions than county–once you get below $60,000 in annual income there is an increase in pollution.





Implications of Findings

MEANING

Can we utilize economic data to improve social change?

Is there merit to uncovering findings that can be visualized to take one step farther away from the **perceived bias of a “We” versus “Them” mentality** that Hans Rosling introduced and meaningfully visualized.

How else can we leverage existing datasets to better illuminate concentrated areas of mortality rates, low household income and a deadly proximity to industrial plants, than to demonstrate that there are literally areas of America that you do not want your family to live.

Politics aside, climate change is a hot-button issue and is deserving of a data-driven, agnostic approach that can re-frame the conversation in a more immediate way that reveals the impact on real individuals.

POLLUTION AND POLICYMAKERS

Pollution is a common byproduct of economic activity. Although policymakers should account for both the benefits and the negative externalities of polluting activities, it is difficult to identify those who are harmed and those who benefit from them. To overcome this challenge, our approach combines datasets that include state size, quantity of industrial plants, type of industrial plant, housing prices, gender, mortality rates, and concentration of pollution.

As reference, the mid-20th century expansion of the U.S. power grid highlights the importance of considering both the costs and benefits of polluting activities, and suggest that **demand for policy intervention may emerge only when the negative externalities are significantly larger than the perceived benefits.**

With Additional Time

Team Perfect Ten's goal would be to continue analyzing reputable datasets and derive a deeper understanding of every state in the union so as to arrive at an accurate matrix similar to the breakdown that we demonstrated for the state of Wisconsin.

For example; Dataset provided county size criteria variation is too dramatic, would like to pursue lat, lon driven mapping/plotting.

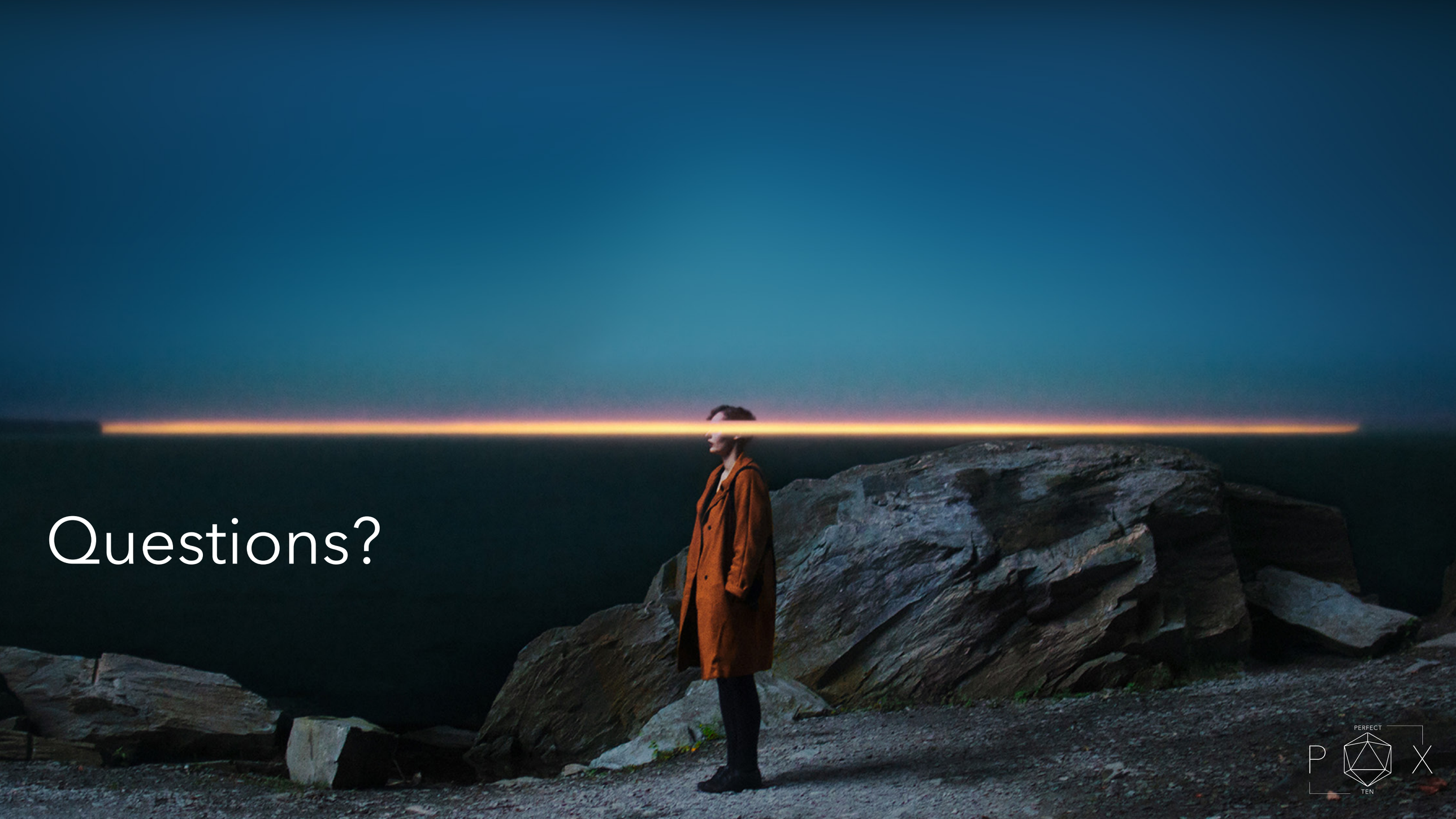
Men, women, life expectancies by both income and pollution is that women were living longer.

Consideration Set

- Grouping by facility type and doing analysis on the type of facility for example processing plants, dairy, paper mill, power plants
- Address change, (spend whole life growing up next to plant–grow up and make more income)
- Pursue more granular outcomes that include lat, lon geo location versus the existing break out by county
- Compare industrial plant type to pollution amount by county, state, country
- Apply similar efforts to other densely populated countries that are already experiencing wide-scale pollution issues, China, India, ect...


“The differences are
much bigger than the
weaknesses of the data.”
- Hans Rosling,
(Pioneer of gapminder.com)
PERFECT
P X
TEN


Thank You


Questions?