Marissa Maki Gill
Env 154: Lab 7

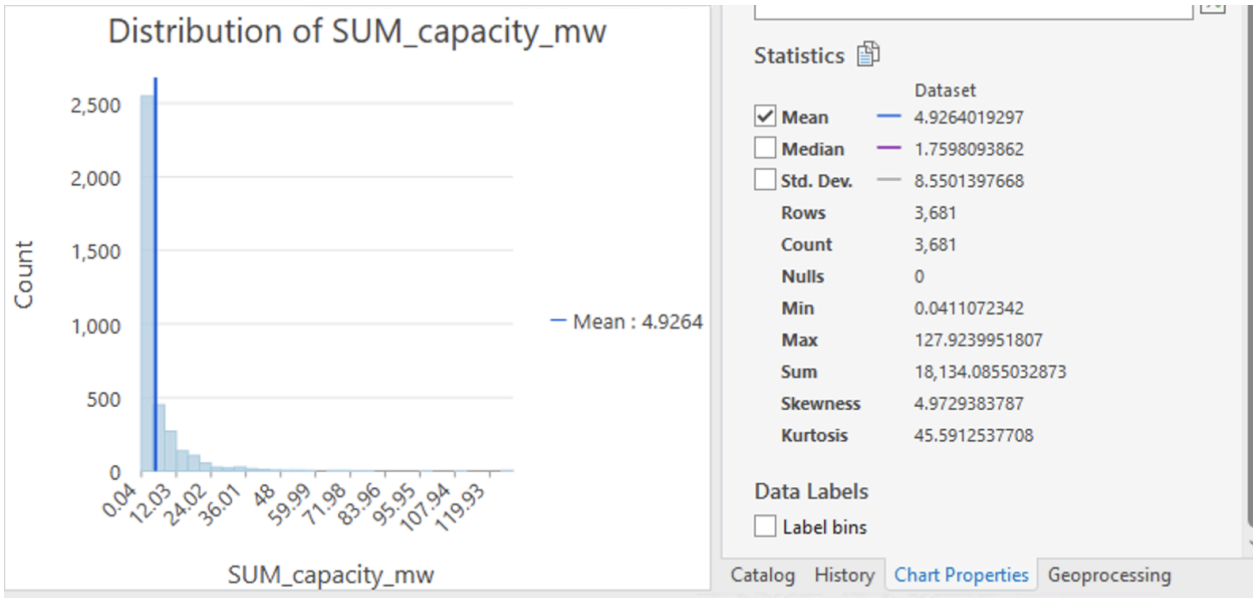
Q1) How many census tracts in the US have at least 1 solar installation?

3760

Q2) How many data points (rows) did you lose as a result of the data clean up (removing Null values and No Values)?

79

Q3) Take a screenshot of your distribution and paste it here. What type of distribution would best fit the SUM_capacity_mw data? Why?



A Poisson distribution would best fit the data because it measures count data, since sum capacity represents the number of solar installations per census tract then it will best be summarized by this right skewed distribution because some areas will have many installations and others will have none.

Q4) Examine your GLR results table and GLR diagnostics tables

a. Copy and paste your GLR results table and GLR diagnostics table below.

Summary of GLR Results [Model Type: Count (Poisson)]

Variable	Coefficient[a]	StdError	z-Statistic	Probability[b]	VIF[c]
Intercept	2.259210	0.081912	27.581079	0.000000*	--------
EP_POV	-0.009471	0.001172	-8.082187	0.000000*	2.339564
EP_UNEMP	0.017971	0.002160	8.320645	0.000000*	1.545548
EP_PCI	-0.000016	0.000001	-17.018529	0.000000*	2.063603
EP_MINRTY	0.005486	0.000453	12.112084	0.000000*	2.461582
EP_NOHSDP	0.003896	0.001027	3.794551	0.000148*	2.560650
POPDEN_MEAN	-0.001456	0.000026	-55.230183	0.000000*	1.278079
CF_MEAN	-0.094629	0.325394	-0.290814	0.771194	1.371876

GLR Diagnostics

Input Features	solarLoc_albers_SpatialJoin_Dissolve_NotNull	Dependent Variable	SUM_CAPACITY_MW
Number of Observations	3681	Akaike's Information Criterion (AICc)[d]	33055.000000
Average Count	4.902744	Deviance Explained[e]	0.237963
Joint Wald Statistic[f]	7462.289930	Prob(>chi-squared), (7) degrees of freedom	0.000000*

b. Which explanatory variables are significant?

EP_POV, EP_UNEMP, EP_PCI, EP_MINIRTY, EP_NOHSDP, POPDEN_MEAN all display statistical significance because they exhibit p-values < 0.05

c. How would you describe the relationship between EP_UNEMP and capacity of distributed solar?

The positive coefficient of the EP_UNEMP variable tells us that the higher the unemployment the more solar distribution in the area.

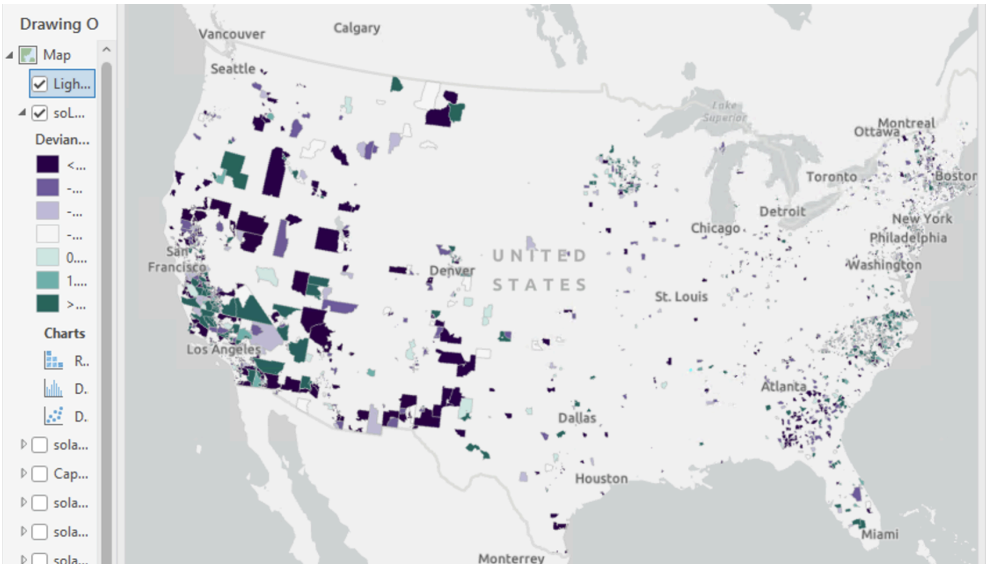
d. How much of the variation in the data is explained by the explanatory variables (Deviance explained, or the R-squared)?

23.83% of variance is explained by the model's deviance explained

e. If one were to use this to predict the capacity at other census blocks, does this seem like a robust model based on the diagnostics (e.g., Join Wald Statistic and deviance explained)?

The low deviance suggests that variation in the solar capacity may be driven by factors not within the model so it may not be reliable for accurate predictions at other census blocks. The join wald statistic is also incredibly high making the model not necessarily statistically significant.

Q5) Copy and paste a screenshot of your residual map below. Can you tell based on just looking at the map of residuals whether there is clustering of residuals?



The residual map displays potential clustering in several states on the East Coast, and Southern California, but it is hard to tell. The map also may show what looks to be slight random distribution in the middle of the states. Just by looking at the map there is clustering of purple in the Georgia/Florida region and clustering of blue in the Virginia/ North Carolina area.

Q6) How do the two Moran's Is and corresponding p-values compare? What do these say about whether the residuals are autocorrelated?

Inverse Distance:

Global Moran's I Summary

Moran's Index	0.087082
Expected Index	-0.000272
Variance	0.000010
z-score	27.202349
p-value	0.000000

K nearest neighbor:

Global Moran's I Summary

Moran's Index	0.162455
Expected Index	-0.000272
Variance	0.000059
z-score	21.139511
p-value	0.000000

Succeeded at Thursday, February 27, 2025 7:42:05 PM (Elapsed Time: 1.35 seconds)

The two p-values from each Moran's I were less than 0.05 indicating significance, and that the residuals were not randomly distributed but that there is spatial autocorrelation between the residuals.

Q7) Review this article on "what they don't tell you about regressions" and think about these suggestions in the context of your regressions results and diagnostics above. What are some things you could do next (more data, more analysis, different parameters) to better answer your question and why?

Some next steps to answering the question include collecting more data and performing more analysis. More socioeconomic data can be collected for analysis like homeownership, housing type or land use as well as political data like meter policies or solar incentives. This data could be helpful because areas with policies promoting solar may have more installations, and geographical data like land use is important to account for because of zoning regulations and proximity to transmission lines. After improving the variety of data another GLR could be run adding the new explanatory variables, and compare the deviance between the original and the updated results. Then another Moran's I could be run to check the spatial autocorrelation and determine which factors now have the strongest impact on solar distribution.

Q8) Finally, sum things up by writing out an answer to the research question we posed at the beginning of this lab assignment so that you can report back to your boss. Include in your answer whether or not each explanatory variable is positively or negatively correlated with higher distributed PV installations and some of the limitations of your model by discussing the results from the Moran's I analysis. As a reminder, you may need to look up the explanation of the SVI variables on the CDC website linked in the first table.

Every positive is positively correlated with higher distribution (list each pos and neg)
Include strength the higher coefficient the more it affects the distribution, and moran's autocorrelation shows spatial correlation, the model may not be capturing everything

The results from the spatial analysis tell us that the socioeconomic and technical factors best correlated with high rates of solar PV adoption at the Census block level in the U.S are EP_MINIRTY, EP_NOHSDP, EP_UNEMP. These explanatory variables showed positive coefficients indicating positive correlation to higher distribution of solar installations. Each variable with negative coefficients (EP_POV, EP_PCI, popDEN_mean, and CF_mean) describes negative correlation to the distribution (no/little correlation). However each variable except the CF_mean has a significant p-value indicating the results are likely not due to random chance alone. The Moran's I that was run also produced statistically significant p-values (<0.05) suggesting clustering of residuals was spatially autocorrelated

and the variation is not due to random distribution. This implies a limitation to the model because our results showed that some areas have higher or lower adoption than predicted by the model, likely due to missing spatial factors that influence the distribution.