CITADEL | CITADEL | Securities

TheDataOpen

Midwest Regional Virtual Datathon Spring 2023

Sania Banga, Siya Digra, Alex Jian
Team 11

February 12, 2023

1 Topic Question

Greenspaces are essential components of our lives, known to improve mental health, cool surface temperature, and reduce noise. However, although greenspaces are so well-known for being beneficial, there still exist neighborhoods throughout the United States in which greenspace access is limited. In this report, we will leverage datasets on tree cover percentages in urban areas, income levels in counties throughout the United States, and the prevalence of poor mental health in different regions to examine the following questions:

1. Is there inequality in access to greenspaces between higher and lower income areas? We will specifically focus on parks, which are supposed to be publicly funded and thereby equally prioritized regardless of area. Can we use this to show discrimination against certain areas in cities?

2. Are the benefits provided by greenspace for mental/physical health more significant for lower income areas than higher income areas? By extension, can additional spending on greenspaces for low income areas deliver beneficial results?

3. Can increased spending on greenspaces in low-income areas contribute to gentrification and displacement of low-income communities? Additionally, if there is evidence of gentrification, how can we change the way greenspace is introduced to prevent this?

Through the answers to these questions, we hope to guide policymakers toward funding better greenspaces in such a way that benefits the mental health of residents of those communities as well as preventing any displacement or gentrification that may occur. With these goals in mind, we will conduct hypothesis tests to answer each of our questions and provide insight into both mental health and inequality as they relate to greenspace.

2 Executive Summary

We begin with an exploratory analysis of depression and mental health in the United States by tract, as well as observing tree cover and mean household income by tract. Green spaces are found to be highly correlated with mental health. Green space provides scientifically proven health benefits by filtering air, removing pollution, attenuating noise, cooling temperatures, infiltrating storm water, and replenishing groundwater; it can also provide food.

We also find that green space is not always equitably distributed, it is often highly stratified based on income. Lower access to greenspace and lower quality of greenspace is positively associated with lower socioeconomic status. Low income groups lack access to other health-promoting sites, such as gyms, rec-centers, etc. so they need greenspaces more, hence greenspaces are extremely necessary for them. Wealthier households often reside on the suburban outskirts where green space is abundant, well-serviced, and well-maintained. This environmental injustice must be taken into account to ensure the benefits are utilized by all communities.

However, addressing park poverty in low-income households may result in a green space paradox. More green space can enhance attractiveness and public health, making neighbourhoods more desirable. As a result, housing costs may rise. Such increases in housing costs have the potential to result in gentrification: the displacement and/or exclusion of the very residents that the green space was intended to benefit. As a result, residents may face higher rents and become homeless, while those who are actually displaced may be forced to leave their communities, ending up in less desirable neighbourhoods with similar park-poverty issues. This paradox has negative public health consequences, not only because of continued park poverty, but also because displacement and precarious housing status have negative public health consequences.[4]

In addition to the descriptive analysis presented above, our modelling work sheds more light on the relationship between mental health, green space, and household income. We built a Linear Regression Model as well as a Random Forest Regression Model, and discovered that the results were consistent with our intuition. Setting up Depression as the Dependent variable and park cover, Income percent as the independent variables/ predictors, we find that depression was negatively linked with green space, as well as household income: Households with lower access to green space and low-income household tend to suffer more from depression. We found the root mean squared error to be 0.21, indicating that our model can relatively predict the data accurately.

3 Technical Expositions

3.1 Data cleaning and pre-processing

3.1.1 Datasets from the Datathon Package

We used the datasets provided to us, specifically the `urban_tree_canopy.csv` and the `PLACES_2022_zcta.csv`. These datasets contain a massive amount of data that was not initially in the format required for our modelling purposes.

To measure the park cover,we extracted the percent park coverage from `percent_cover_tracts_with_buffer.csv`. We note that these park coverages were measured in 2019, however, since park coverage is relatively constant over a three year period (our other data is from 2020-2022), we assumed that these measurements would still be relatively accurate for the near future. We then started to analyze `urban_tree_canopy.csv`, and extracted the fields that would help us analyze the relationship between park cover and mental health: census block, mean percent tree cover, tree gap and income percent.

From the `PLACES_2022_zcta.csv` dataset we extracted the tract and only the depression and mental heath illnesses from the Data Value field, since we were only focused on mental health. To make these datasets more comparable, we joined them on tract and began our data analysis.

3.1.2 External Datasets

To explore the effect of greenspace in low-income areas on gentrification, we collected yearly park spending budget dataset from `'https://www.tpl.org/parkserve'`. We draw the City based Spending per Resident for Parks for the years 2016 to 2019. To analyse the effect, we also collected the cost of living index for cities from `'https://www.numbeo.com/cost-of-living/region_rankings'` for the years 2016-2019. The purpose behind analysing the Spending per resident for parks and cost of living index of the cities over the years is to see if there is an increase in cost of living index for gentrifiable cities, that is, cities with low spending per resident in 2016-2017.

3.2 Exploratory data analysis

3.2.1 Mental health prevalence by zip code

In order to begin understanding the data for mental health in the United States, we create an initial graphic to examine the differences in the prevalence of poor mental health in different zip codes in the United States. We used the provided `PLACES_2022_zcta.csv` dataset, which contains `Data_Value`, the estimated percent prevalence of a certain disease, disorder, or condition among a US zip code, which is calculated using a linear regression model formulated by the CDC. Each row of the dataset contains several defining values for the location, category, measure, and data value as described above. When parsing the dataset, we specified the data type of `'LocationName'` to be of type `str` to preserve the leading zero in the zip code.

We now filter the dataset to only contain rows with the values `'MHLTH'` and `'DEPRESSION'` for the column `'MeasureId'`, which is the key describing the type of condition. The former value denotes poor general mental health, while the latter is individuals diagnosed with a form of depression specifically. Note that the aforementioned values are the only two values in the dataset denoting mental health-related disorders or conditions. An initial look at the data shows us that the distribution of mental health prevalence across zip codes is approximately bell-shaped, as both the mean and median of `Data_Value` fall at around 18%. The maximum prevalence of a mental health condition is around 40%, while the minimum lies around 4%. The standard deviation of the prevalence is roughly 4.4.

With no immediate skew visible, our next step is to create the visualization. Using a shape file for US counties, we created a geographic visualization for both the `'MHLTH'` value and the `'DEPRESSION'` value.

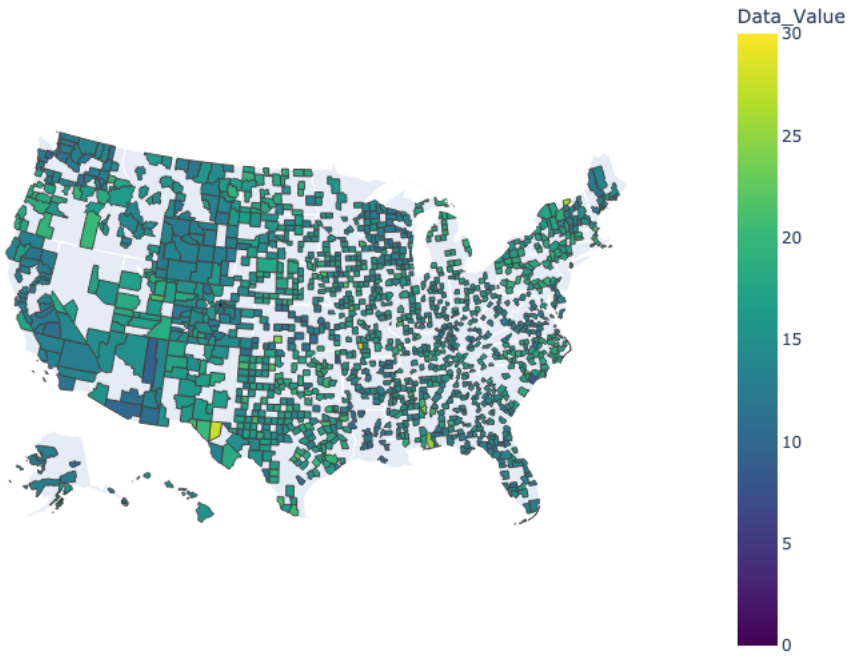

Figure 1: Percent prevalence for poor mental health by zip code

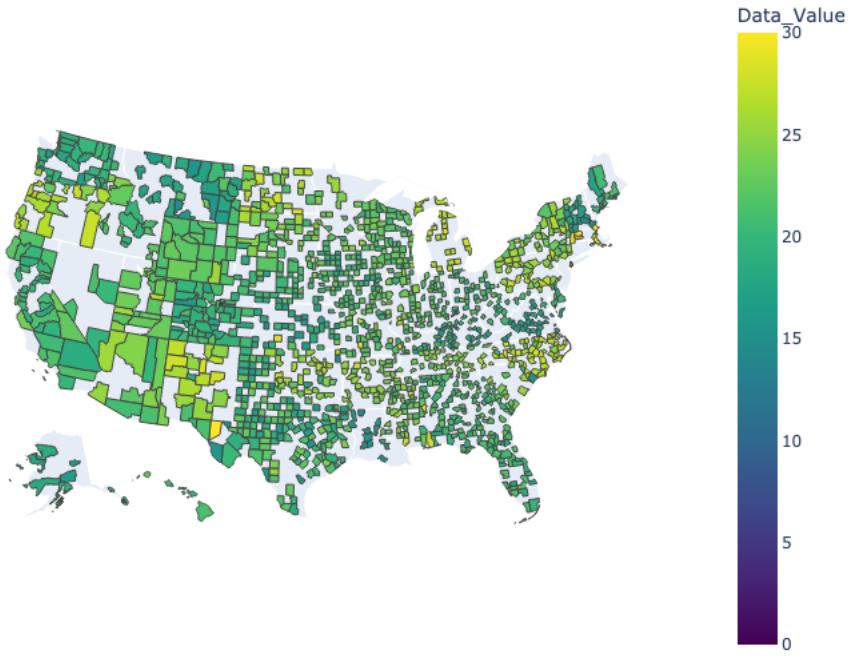

Figure 2: Percent prevalence for depression by zip code

An initial look at this may seem to reveal a spike in poor mental health and depression around the Southwest region. It is important to note that there were the same number of datapoints for both types of measures (around 32,000), so there exist no discrepancies in terms of the amount of data for either. However, it is apparent that data is missing for many zip codes and thus, we cannot get a comprehensive view on mental health throughout all US zip codes. We also observe that Fig. 1 shows lower prevalence in general

than Fig. 2. This may be because the poor mental health measure does not account for individuals with specific disorders, and rather those who have poor mental health without a concrete diagnosis. Thus, this may explain why the prevalence for any particular disorder may be represented as higher even though they all fall into the category of poor mental health.

3.3 Modeling

We investigate each of the questions posed above primarily by using linear regression models and the corresponding hypothesis tests. First, we establish a link between tree/park coverage and depression scores, independent of other factors such as income, population density, or surface temperature. We then demonstrate a clear correlation between income levels and greenspace coverage at the census tract level, showing that lower income areas suffer from lower access to parks and green areas, despite the fact that parks are publicly funded and supposed to be equally distributed among all areas. We examine the difference between the marginal benefit in terms of mental health of a unit park coverage percentage increase for low income versus high income areas, allowing us to see whether greenspace initiatives would be especially helpful in low income areas. Finally, we investigate possible side effects of greening initiatives, namely, gentrification, in lower income neighborhoods and cities.

3.3.1 Park Coverage and Mental Health/Depression

For the link between tree/park coverage and mental health/depression scores, we first establish a linear regression model linking all our available data to depression scores. The regression is:

$$dpmh = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \beta_4 * x_4 + \epsilon \tag{1}$$

Where dpmh is the depression prevalence/mental health score, β_0 is the intercept term, x1 is the average annual income for the given census tract, x2 is the tract's population density, x3 is the average surface temperature, and x4 is the log of the percentage park cover for the tract.

We first report R2 values to check the correlations between our predictors, verifying that since none are above 0.1, we should not have any serious collinearity issues. We also check the distributions of our predictors themselves, noting that the percentage park cover data was heavily skewed right:

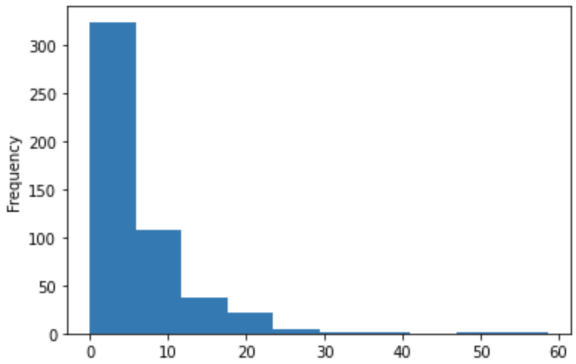

Figure 3: Distribution of park coverage percentages for census tracts

As a result, we normalize by log-transforming this predictor. Then, we run an OLS linear regression on all variables, combined with significance tests, and report the results in the table below:

OLS Regression Results

Dep. Variable:	y	R-squared:	0.129
Model:	OLS	Adj. R-squared:	0.129
Method:	Least Squares	F-statistic:	5.877e+04
Date:	Sun, 12 Feb 2023	Prob (F-statistic):	0.00
Time:	20:19:26	Log-Likelihood:	-4.0556e+06
No. Observations:	1588318	AIC:	8.111e+06
Df Residuals:	1588313	BIC:	8.111e+06
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	21.3104	0.024	883.619	0.000	21.263	21.358
x1	-5.353e-05	1.37e-07	-391.771	0.000	-5.38e-05	-5.33e-05
x2	0.0676	0.001	84.807	0.000	0.066	0.069
x3	-0.5520	0.003	-195.937	0.000	-0.558	-0.547
x4	-0.3328	0.003	-113.374	0.000	-0.339	-0.327

Omnibus:	11125.592	Durbin-Watson:	0.026
Prob(Omnibus):	0.000	Jarque-Bera (JB):	10523.813
Skew:	0.170	Prob(JB):	0.00
Kurtosis:	2.793	Cond. No.	3.84e+05

Figure 4: Linear Regression Coefficients and Significance Tests

The table shows the results of the significance test for β_4, the regression coefficient associated with percent park cover (x4). The significance test is a t-test conducted on the null hypothesis that beta4 = 0. Since our model already includes the other variables that impact depression prevalence, this t-test tests whether x4 is correlated with dpmh, *conditioned on x1, x2, and x3*, which intuitively means whether x4 has a significant effect on dpmh after accounting for the effects of the other variables.

Our t-test shows an extremely large t-value (-113.374) and a p-value close to 0, showing that park coverage does have a significant correlation with depression prevalence and mental health, even after accounting for effects explained by differences in income, population density, or surface temperature. While this still does not definitively prove a causation between park coverage and depression/mental health, it gives strong evidence that the two are related, and the negative regression coefficient (-0.3328) shows that increasing park space coverage can decrease mental health issues.

3.3.2 Income and Tree/Park Coverage

With the link between tree/park coverage and mental health established, we now turn towards potential inequality in park accessibility/availability in lower income areas. We begin by simply plotting the correlation between income and park coverage, per census tract. Note that we threw out census tracts with very high income (¿150000) to prevent high leverage points (points whose x-value deviates far from the mean, and carry a high influence on correlation and regression) from influencing our visualization too much.

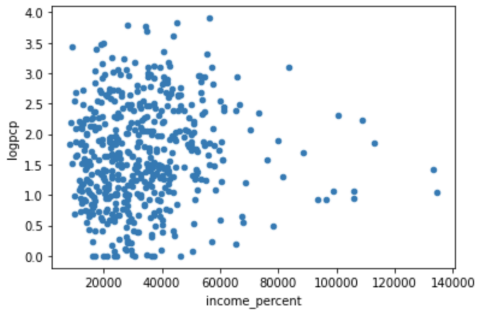

Figure 5: Income vs Park Cover

We see a weak, but definitively clear positive correlation in the graph, showing that lower income areas likely have less access to parks than higher income areas. We then perform an independent t-test for low

income areas (less than 35000 dollars annual income) against higher income areas (greater 35000 dollars annual income) to analyze this difference in park cover more rigorously. We obtain:

Group 1 (Low income areas): Mean = 0.274, Std = 0.205
Group 2 (Mid to High income areas): Mean = 0.338, Std = 0.224

Conducting an independent t-test on the data gives a t-value of -206 and a p-value of very close to 0, showing a significant different in park coverage between low and mid to high income areas. While this is not enough to show discrimination, we can definitively conclude that lower income areas (census tracts) have, on average, less access to park space and green areas. Since we have demonstrated a link between greenspace access and mental health, this finding shows a potential contributor to mental health problems in low income areas and a potential need to equalize greenspace and park access, especially since it is a publicly funded amenity that should be equally accessible regardless of income or area.

3.3.3 Gentrification

Urban greening initiatives can cause gentrification, a process whereby affluent residents are attractive to previously lower income areas, driving up living costs and uprooting the original long-term residents of the area.

To investigate this phenomenon, we looked for large jumps in city park spending between consecutive years in our yearly park spending dataset, which likely indicated urban greening initiatives beginning. Then, in low-income census tracts belonging to these cities, we looked for a rising cost of living index over subsequent years, indicating a possible gentrification process taking place. However, this exploration did not pass the exploratory phase: we found that the cost of living index did not increase significantly during subsequent years for areas with large jumps in greenspace/park spending.

4 Open Questions

While we were unable to determine a causal relationship between greenspace and gentrification, we may be able to deliver more meaningful results with park spending data that includes more cities/counties. Along with that, the result of gentrification could be more easily shown if we had access to demographic data across cities/counties that is measured annually rather than every decade.

We could also build a better model to predict gentrification with if new variables are taken account for in the model. At the moment, we are leveraging rental prices, per resident income, and demographic composition. We could also use home values (instead of only rental prices), and the percent of buildings built after a particular year, which could be a good indicator for new residential buildings built due to gentrification.

Additionally, in terms of mental health, we were given prevalence values measured by zip code, county, tract, and census-designated places. Perhaps analyzing the predicted prevalence values in all four of these different regions may yield different or more specific results for the correlation between greenspace and mental health. As noted before, the data for all regions in the `PLACES_2022_zcta.csv` were not given, so perhaps a stronger causal inference could also be derived between tree cover and mental health prevalence while accounting for varied income values in those areas.

References

[1] Barton J., Rogerson M., 2017, BJPsych International
[2] Hoffimann, E., Barros, H., Ribeiro, A.I., 2017, Int Journal of Environmental Research and Public Health
[3] Black, K. J., Richards, M., 2020, Landscape and Urban Planning
[4] Wolch R. J., Byrne, J., Newell, J. P., 2014, Landscape and Urban Planning