CS51 - Correlation and Regression

Minerva University

CS51: Formal Analyses

Prof. Volkan

January 29, 2023

Correlation and Regression Report - Historical Racism Redlining, Intra-Urban Heat, and Tree Cover Data

I. Introduction

In the 1930s, to prevent foreclosures[1] and increase the affordability of loans and homeownership post-Great Depression, the Home Owners' Loan Corporations (HOLC), as part of federal programs, created residential maps of 239 individual US cities to determine their credibility and security for lending and real-estate investments (Swope et al., 2022). This led to a discriminatory practice called "redlining" which ranks the loan worthiness of those areas based on their racial characteristics. Wealthier neighborhoods were color-coded as green and blue, while underprivileged neighborhoods were labeled yellow and red.

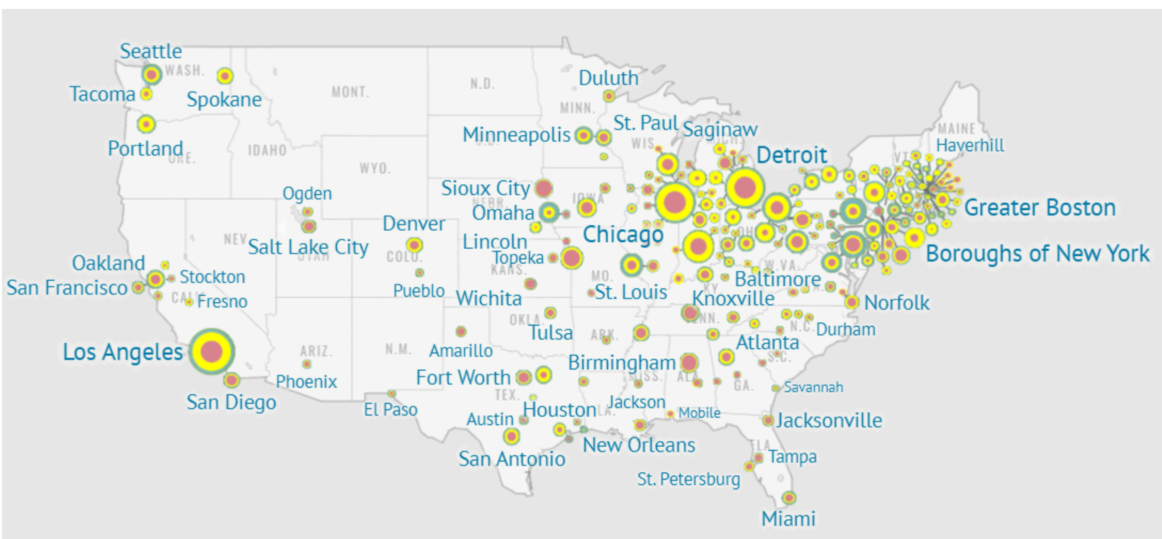


Figure 1. [Mapping Inequality - Redlining in New Deal America](). The size of each circle represents the area in that city that HOLC graded, with each color representing the proportion of the city graded and colored. Green (A) - Best, Blue (B) - Still Desirable, Yellow (C) - Definitely Declining, and Red (D) - Hazardous.

---

[1] The legal process by which a lender seizes/sells a home or property after a borrower is unable to meet their repayment obligation.

Communities of color and low-income households have been bearing severe consequences due to such historical practice, particularly the alarming heat-related public health (Li et al., 2022). Thus, it is worth studying the relationship between the land surface temperature and the percentage of tree cover in D regions to inform policymakers and non-profit organizations about the most vulnerable regions to target. This report applies regression and statistical significance methods to evaluate the relationship.

II. Dataset

This dataset is extracted from the research paper "The Effects of Historical Housing Policies on Resident Exposure to Intra-Urban Heat: A Study of 108 US Urban Areas" [2]. Four missing values of tree cover percentages in Lake Country and one missing value of A Tree cover % in G.NYC Area are removed before the analysis.

	Lansat Date	Urban area	State	A δLST	B δLST	C δLST	D δLST	D-A (°C)	A Tree cover %	B Tree cover %	C Tree cover %	D Tree cover %
0	29-Jul-17	Joliet	IL	0.70	0.95	-0.10	-0.77	-1.47	15.538567	11.793579	14.338625	16.733715
1	9-Aug-17	Lima	OH	2.64	-2.07	0.08	1.81	-0.83	29.630212	17.495285	18.889741	15.980380
3	1-Aug-14	Pontiac	MI	1.07	-0.15	-0.58	0.68	-0.39	25.732175	30.573424	16.246246	17.307430
4	20-Jun-17	Evansville	IN	-0.08	0.02	0.28	-0.47	-0.39	21.608707	16.392225	16.892633	18.964544
5	5-Jun-14	Saginaw	MI	0.04	-0.16	0.06	-0.10	-0.14	26.461557	22.847605	15.906192	14.476868
...	...	...	...	...	...	...	...	...	...	...	...	...
103	14-Jun-17	San Francisco	CA	-2.16	-1.04	1.02	1.93	4.09	17.488805	8.807723	8.022475	7.156359
104	7-Jun-17	Fresno	CA	-3.49	-0.54	0.40	0.61	4.10	13.358519	8.557843	4.634166	4.081426
105	9-Jun-17	Los Angeles	CA	-3.03	-0.56	0.99	1.18	4.21	13.206908	8.505291	5.675662	4.522813
106	1-Aug-16	Denver	CO	-4.09	-2.08	0.40	2.59	6.68	22.215545	14.253873	9.137111	5.485505
107	28-Aug-16	Portland	OR	-4.42	0.52	0.72	2.67	7.09	45.849071	23.685191	16.072903	15.513399

Table 1. Redlining and Climate Change Data Frame. The sample size n = 106 after removing two rows of areas that have missing values for tree cover %

[2] The researchers condense the 239 unique HOLC maps into a database of 108 US cities or urban areas that overlap within Landsat 8 imagery tiles, and excluding any cities that were not mapped with at least one of all four HOLC security rating categories (n = 4). U.S Census Bureau regions: Northeast (n = 26), South (n = 29), Midwest (n = 41), and Western (n = 12).

All variables used for this report are quantitative continuous - measured numerically and taken infinite values in the form of decimals with no defining pairs of consecutive values, which is appropriate for scatter plots, histograms, and regressions. The two tables below provide more detailed information about the chosen variables, the research question, and the hypotheses guiding this report.

Variable name	Variable type	Description
D Tree cover (%)	Predictor (x)	The data of tree cover percentages (or tree canopy) are extracted from the National Land Cover Database (NLCD) 2011. NLCD imperviousness reports the percentage of urban developed surfaces that are not affected by heat over every 30 m pixel in the United States.
D δLST (°C)	Response (y)	**δLST (Land Surface Temperature Anomaly)** estimate shows relatively how much warmer or cooler a particular HOLC security rating polygon is from the entire set of HOLC security rating polygons for a given urban area, and then compares these anomalies between cities in a quantitative manner. Polygon basically means a filled region with clear boundaries on a map. Below is the formula to calculate the anomaly: $\delta LST_{area,\ polyglon} = \overline{LST_{area,\ polyglon}} - \overline{LST_{area,\ all\ polyglons}}$

Research Question	Is the tree cover of a given city in D region a good predictor for the land surface temperature anomaly there?
Null Hypothesis	H_0: The percentage of tree cover in a D city cannot predict the temperature anomaly there. ($\beta_1 = 0$)
Alternative Hypothesis	H_A: The percentage of tree cover in a D city can predict the temperature anomaly there. ($\beta_1 \neq 0$)
Hypothesis Testing	Type I Error: D Tree cover cannot predict the temperature anomaly, but I conclude that it can. Type II Error: D Tree cover can predict the temperature anomaly, but I conclude that it cannot. [3]

III. Methods & Results

Summary Statistics	Predictor Variable (D Tree cover)	Response Variable (D δLST)
Count	n = 106	n = 106
Mean	$\bar{x} = 15.845$	$\bar{y} = 0.7896$
Standard deviations	$s_x = 8.046$	$s_y = 0.928$

[3] **#variables:** As the initial step to examine the relationship between temperature anomaly and tree cover, two variables, and their roles have been identified and classified with relevant information where they are extracted from and how they are calculated. The type of variables is also clearly stated at the beginning (quantitative and continuous) to prove that they are appropriate for the regression analysis. Particularly, the relationship between them is suggested in the second table to guide later calculations and testing.

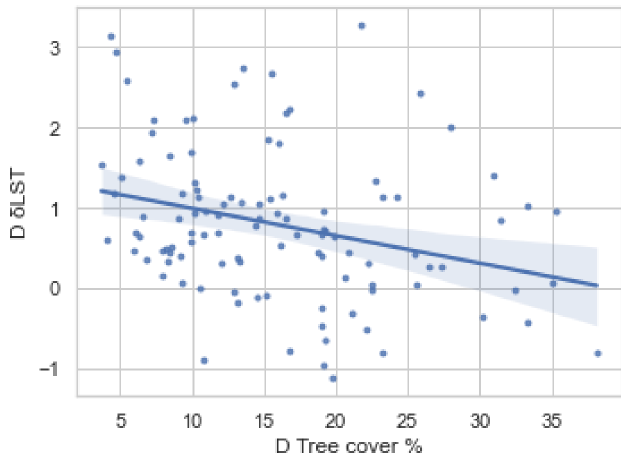


Figure 2. A scatter plot of tree cover percentage (x-axis) against the land surface temperature anomaly (y-axis) of individual cities in the 'D' neighborhood.

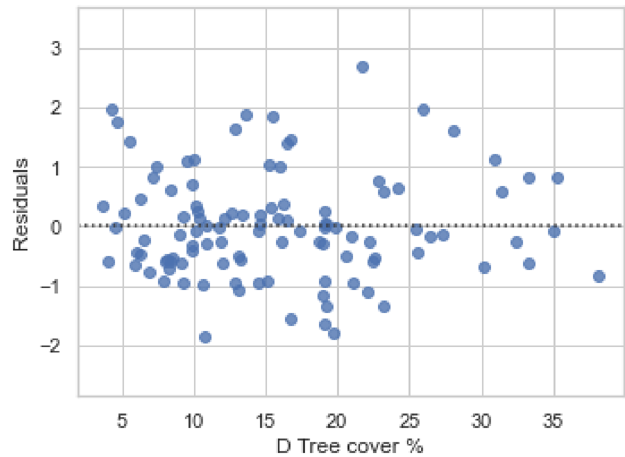


Figure 3. A scatter plot showing the fitted values of D tree cover on the x-axis and the residuals (actual value - expected value).

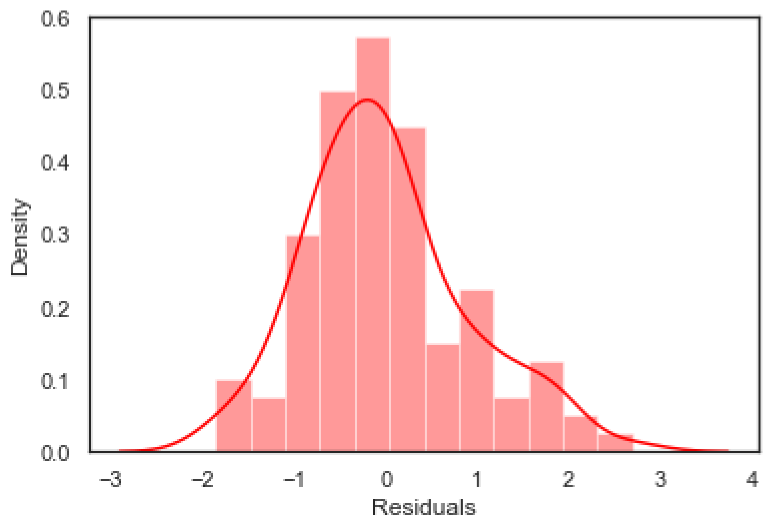


Figure 4. A histogram of the residual distribution with a normal bell curve[4]

1. Conditions Check (LINER)

[4] **#dataviz**: I have drawn three diagrams (scatterplot for x-variable and y-variable, residual scatterplot, and residual histogram) for this dataset because they are the most useful type of visualization to check the conditions for this report. The diagrams are appropriately labeled and a bell curve is included to show the normality of the residual in the histogram. Explanations are followed to provide information about the sample.

Firstly, figure 2 illustrates the approximately linear relationship between two variables because it is football-shaped. Most of the data points fall further from the regression line, meaning that the difference between the actual and expected values is large. Secondly, the independence of errors is also guaranteed as figure 3 shows their random scatteredness with little influence on each other. Thirdly, figure 4 demonstrates the nearly normal distribution of the residuals, even though it is slightly right-skewed. The variance of the residuals is tube-shaped and strongly equal within the range of -2 and 2 (homoscedasticity). Also, there are over 3500 urban areas in the U.S. (Bureau, n.d.), and the sample size is 108, which satisfies the ≤ 10% population rule for independence. Finally, it is assumed that the data from the cities were collected randomly. Given that all the conditions are met, the regression procedure can be conducted.

2. **Correlation Coefficient (r)**

$$r = \frac{1}{n-1}\Sigma\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

n (sample size) - x_i, y_i (specific values of x and y variables) - $\bar{x}$, $\bar{y}$ (means of all x-variables and y-variables) - s_x, s_y (standard deviations of x-variables and y-variables).

Pearson's r value comparing D Tree cover % to D δLST is r = - 0.291. Since the correlation coefficient is moderately negative, it is inferred that the bivariate data are linearly correlated and downwardly directed - when the value of x increases, the value of y decreases. Since r is very close to 0, the strength of the correlation is weak. It is noteworthy that this does not imply causation between the two variables because of some reasons that will be addressed in the conclusion. [5]

---

[5] **#correlation:** the assumptions needed for regression including linearity, independence, normality, equal variance, and randomness are checked before proceeding to the analysis to make sure the results are reliable. The correlation between temperature anomaly and tree cover was calculated and the r-value suggests that a linear relationship exists. These results are also interpreted fully to reach a conclusion about the strength of the correlation and the predictability of y based on x, which is not strong.

3. Coefficient of Determination (R^2)

$$R^2 = 1 - \frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - \bar{y}_i)^2}\ (\hat{y}\ is\ the\ predicted\ value) = 1 - \frac{SSE\ (variation\ explained\ by\ the\ model)}{SST\ (total\ variation)}$$

$$b_1(slope) = \frac{\Sigma y \Sigma x^2 - \Sigma x \Sigma xy}{n(\Sigma x^2) - (\Sigma x)^2}\ ;\ b_0(intercept) = \frac{n(\Sigma xy) - \Sigma x \Sigma y}{n(\Sigma x^2) - (\Sigma x)^2}$$

$R^2 =$ 0.084 | Regression equation: $\hat{y}\ (D\ \delta LST) = -0.034x\ (D\ Tree\ cover) + 1.32.$

To determine how well the model fits the data, we divide the (squares of the difference between y actual values − y predicted values) by the (squares of y actual values − y mean). As R-squared is relatively low, the sum of squared errors - the difference between observed and predicted values - is high, meaning the data points scatter widely around the fitted regression line. Only 8.4% of the variability in the temperature anomalies of cities in hazardous D areas can be explained by the percentage of tree cover there. Thus, the model does not fit the data very well.

The slope parameter is - 0.034, indicating a negative correlation between the two variables. More specifically, with every additional percentage point increase in tree cover, the average temperature anomalies decrease by 0.034 Celcius degrees. The intercept means that the average temperature anomaly is 1.32 when the tree cover is 0. In this case, it does not make sense to interpret because no cities have zero trees. [6]

4. Hypothesis Testing (p-value)

[6] **#regression:** the coefficient of determination for the simple regression model of D temperature anomaly and tree cover is calculated to conclude how well the model fits the data and probability in which the variability in dependent variable y can be explained by the independent variable x. The slope and intercept of the regression equation are also interpreted to elaborate on the numerical relationship between x and y.

The consequences of committing Type II error, concluding that there exists no relationship between temperature anomalies and tree cover while it actually does, are more serious since the heat-related health issues and deaths in D areas would persist if no interventions are made. Also, as the relationship direction between x and y is uncertain, we conduct a two-tailed hypothesis testing with a significance level (α) of 0.1. Using the formula of t-distribution with n − 2 degrees of freedom (df), the obtained result of p-value is 0.0025. In other words, there is only a 0.25 % chance that the true coefficient of D Tree cover equals 0. Based on this p-value ($p < \alpha$), it indicates there is sufficient evidence in the sample to conclude that the independent variable (D Tree cover %) can be a predictor for the dependent variable (D δLST), in favor of the alternative hypothesis. This result about the relationship between D Tree cover and D land surface temperature anomaly is, therefore, statistically significant.[7]

5. **Confidence Intervals**

$$\mathrm{SE}(b_1) = \frac{s_y}{s_x}\sqrt{\frac{1-R^2}{n-2}}$$

Using the formula $[b_1 \pm t_{df=n-2}. SE(b_1)]$ with $t_{df=n-2}$ is the critical value of the t-distribution with n-2 degrees of freedom and the confidence level of 90% ($\alpha = 0.1$), the obtained confidence interval for the slope estimate is [-0.05, -0.02]. Thus, we can be 90% confident that the true population slope coefficient is within the interval [-0.05, -0.02], indicating that the relationship between tree canopy and temperature anomaly in a given D city is negatively correlated. If we repeated the sampling procedure many times, obtaining different samples and new confidence

[7] **#significance:** a measure of significance was interpreted to explain the observed statistical difference considering the sample size. The justification for the 0.1 significance level is provided, followed by the calculation of the p-value. The p-value is then compared with the alpha to determine whether the relationship between x and y is statistically significant.

intervals each time, 90% of these confidence intervals would capture the true value of the population slope that is different from $\beta_1 = 0$. In other words, the probability of favoring the alternative hypothesis is extremely high. The result is consistent with the p-value testing above. [8]

6. Forward Selection

The starting model has one predictor as D Tree cover: $R^2_{adj} = 0.084$. Then, each of the possible models is fitted with only one variable in the first step. The table below summarizes the adjusted R^2 of 7 models with each predictor:

D Tree cover %	A δLST	B δLST	C δLST	A Tree cover %	B Tree cover %	C Tree cover %
R^2_{adj}	0.148	0.298	0.078	0.225	0.291	**0.299**

The one-predictor model with the largest adjusted R-squared is the D Tree cover % = 0.299. We will add this variable to the model. Then, we continue this process with two-predictor models in which one of the predictors is D Tree cover and the new baseline $R^2_{adj} = 0.299$.

Add C Tree cover %	A δLST	B δLST	C δLST	A Tree cover %	B Tree cover %
R^2_{adj}	0.339	**0.411**	0.320	0.314	0.317

8 **#confidenceintervals:** The population slope is estimated using confidence intervals: the temperature anomaly in D city based on a sample plus/minus the margin of error. The confidence level of 90% is determined accordingly based on the significance level. It is interpreted in terms of the probability that can be inferred for the population. This is used to inform better interventions, noting the implications for statistical significance, consistent with the p-value test.

The two-predictor model with the largest adjusted R-squared is the C Tree cover % = 0.411 (> 0.277). We will add this variable to the model with a new baseline $R^2_{adj} = 0.411$ and repeat the process as three-predictor models. The same applies to the four-predictor and five-predictor models. We choose the largest adjusted R-squared compared to the baseline because it contributes the most to the improvement of the model, and we stop when the addition of new variables decreases the R^2_{adj}

Add B δLST	A δLST	C δLST	A Tree cover %	B Tree cover %
R^2_{adj}	**0.435**	0.417	0.411	0.405

Add A δLST	C δLST	A Tree cover %	B Tree cover %
R^2_{adj}	0.430	0.430	0.430

As a result, we choose 2 additional variables besides D Tree cover %: C Tree cover % and B δLST to add into the model. However, as shown in the correlation matrix plot in figure 5, C Tree cover % and D Tree cover % are collinear, which will reduce the precision for estimates of the regression coefficients. Thus, we will not add C Tree cover % to the regression model. The final multivariate regression includes two predictors (D Tree cover % and B δLST) against the response variable (D δLST).

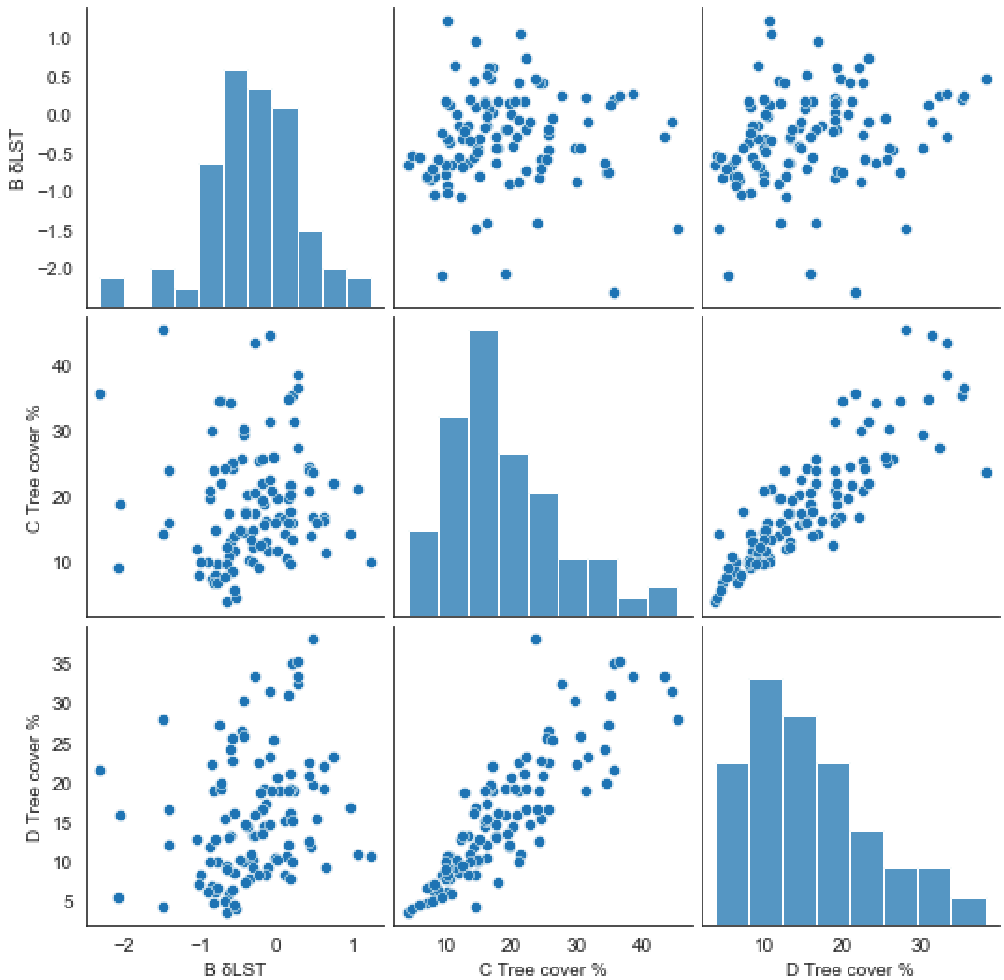


Figure 5. Correlation plots of pairwise relationships between three chosen variables within the dataset.

IV. Conclusions[9]

Based on the correlation coefficient results (r = - 0.291), there is a moderately weak negative correlation between D land surface temperature anomalies and D tree cover. When the tree canopy percentage in D cities increases, the land surface temperature anomaly decreases. The coefficient of determination ($R^2 = 0.084$) infers that only 8.4% of the variability in the D land surface temperature can be explained by the D tree cover, which means that x is not a highly precise predictor for y. However, given the p-value of 0.0025 compared to the significance level of 0.1 and the 90% confidence intervals capturing the true value of the population slope that contains negative values ($\neq 0$), the relationship between D tree cover and D land surface temperature anomaly is statistically significant. Therefore, the answer to the research question is that despite the weak negative correlation between temperature anomalies and tree cover in redlined areas, there is sufficient evidence that the temperature anomaly in a hazardous-labeled city can be predicted by the percentage of tree canopy there ($\beta_1 \neq 0$). The result from this sample can be generalized for the population of over 3500 urban areas in the US. Based on the above premises, such an inductive conclusion is strong.

Nonetheless, many confounding variables not included in the dataset might affect both the temperature anomaly and tree cover: 1) the extensive use of concrete materials to build roads and pathways in some areas which radiates more heat, 2) data collected during summertime (June-August) when the temperature is high, and 3) the climate conditions specific to each region -

---

[9] **#organization:** I organized the report into 5 main parts including the introduction, dataset, methods & results, conclusions, and reflection. This organization starts with descriptive information about the dataset, variables, and their relationships and follows with a more detailed analysis before reaching the conclusion. Particularly, I utilized tables to present information that makes it easier for the readers to follow and tie the conclusion back to the research question.

some are more arid and conducive to sustain tree cover and reduce heat (Hoffman et al., 2020). There are some regions of the country not included in the dataset, which makes the generalizability of the dataset questionable. These limitations render the above conclusion not sufficiently reliable. Eliminative evidence, such as a multiple regression including B δLST as another predictor, should be conducted to examine the relationship between tree cover and temperature anomaly. Some other factors affecting the temperature shift should also be considered so that proper interventions in those underprivileged regions are made.[10]

Word count: 1800 (not including in-text citations, figure captions, tables including only numbers, and mathematical formulas)

[10] **#induction:** the type of induction for this report is defined as generalizability (sample - population). The strength of the conclusion is evaluated based on the results of regression coefficients, p-value, and confidence intervals (premises). However, some confounding variables that impact both the temperature anomaly and tree cover are limitations, making the induction not highly reliable. Suggestions to conduct more test is mentioned to improve inductive reasoning.

References [11]

Hoffman, J. S., Shandas, V., & Pendleton, N. (2020). The effects of historical housing policies on resident exposure to intra-urban heat: A study of 108 US urban areas. Climate, 8(1), 12. https://doi.org/10.3390/cli8010012

Swope, C. B., Hernández, D., & Cushing, L. J. (2022). The relationship of historical redlining with present-day neighborhood environmental and health outcomes: A scoping review and conceptual model. Journal of Urban Health, 99(6), 959–983. https://doi.org/10.1007/s11524-022-00665-z

Li, D., Newman, G. D., Wilson, B., Zhang, Y., & Brown, R. D. (2022). Modeling the relationships between historical redlining, urban heat, and heat-related emergency department visits: An examination of 11 Texas cities. Environment and Planning B: Urban Analytics and City Science, 49(3), 933–952. https://doi.org/10.1177/23998083211039854

Bureau, U. C. (n.d.). Urban areas facts. Census.Gov. Retrieved January 29, 2023, from https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural/ua-facts.html

[11] **#professionalism:** I asked my friends to help detect mistakes that could be missed in my self-proofreading, reviewing all formatting and grammatical mistakes using Grammarly throughout the writing process, and adhering the field's professional standards contributed to the overall assessment of the proposal. All non-original information was properly attributed to its original source in APA-style to abide by the conventions of the field.

Appendix A - General Dataset and Descriptive Statistics

In [2]:
```
#THE EFFECTS OF HISTORICAL HOUSING POLICIES ON RESIDENT EXPOSURE TO INTRA-URBAN HEAT: A STUDY OF 108 US URBAN AREAS
# Import useful packages
import pandas
import numpy as np
from scipy import stats
import matplotlib.pyplot as plt
import statsmodels.api as statsmodels # useful stats package with regression functions
import seaborn as sns # very nice plotting package

# style settings
sns.set_style("white")

# import and print data
df = pandas.read_csv("https://course-resources.minerva.edu/uploaded_files/mu/00294494-8608/redlining.csv")
data = df.drop(columns=['Region']) #remove this column because the "Nan" string values do not contribute to the analysis
data.dropna() #remove 4 missing values of tree cover in Lake Country & 1 missing value of tree cover G.NYC Area
```

Out[2]:

	Lansat Date	Urban area	State	A δLST	B δLST	C δLST	D δLST	D-A (°C)	A Tree cover %	B Tree cover %	C Tree cover %	D Tree cover %
0	29-Jul-17	Joliet	IL	0.70	0.95	-0.10	-0.77	-1.47	15.538567	11.793579	14.338625	16.733715
1	9-Aug-17	Lima	OH	2.64	-2.07	0.08	1.81	-0.83	29.630212	17.495285	18.889741	15.980380
3	1-Aug-14	Pontiac	MI	1.07	-0.15	-0.58	0.68	-0.39	25.732175	30.573424	16.246246	17.307430
4	20-Jun-17	Evansville	IN	-0.08	0.02	0.28	-0.47	-0.39	21.608707	16.392225	16.892633	18.964544
5	5-Jun-14	Saginaw	MI	0.04	-0.16	0.06	-0.10	-0.14	26.461557	22.847605	15.906192	14.476868
...	...	...	...	...	...	...	...	...	...	...	...	...
103	14-Jun-17	San Francisco	CA	-2.16	-1.04	1.02	1.93	4.09	17.488805	8.807723	8.022475	7.156359
104	7-Jun-17	Fresno	CA	-3.49	-0.54	0.40	0.61	4.10	13.358519	8.557843	4.634166	4.081426
105	9-Jun-17	Los Angeles	CA	-3.03	-0.56	0.99	1.18	4.21	13.206908	8.505291	5.675662	4.522813
106	1-Aug-16	Denver	CO	-4.09	-2.08	0.40	2.59	6.68	22.215545	14.253873	9.137111	5.485505
107	28-Aug-16	Portland	OR	-4.42	0.52	0.72	2.67	7.09	45.849071	23.685191	16.072903	15.513399

106 rows × 12 columns

In [8]:
```
tree_cover_D = data_1['D Tree cover %'] #filter the column to take values only in the D Tree cover %
tree_cover_D.describe() #method to return the description of the numerical data in the DataFrame
```

Out[8]:
```
count    106.000000
mean      15.845348
std        8.046114
min        3.710396
25%        9.645955
50%       14.661997
75%       20.450998
max       38.135919
Name: D Tree cover %, dtype: float64
```

In [9]:
```
temperature_delta = data_1['D δLST'] #filter the column to take values only in the D δLST
temperature_delta.describe()
```

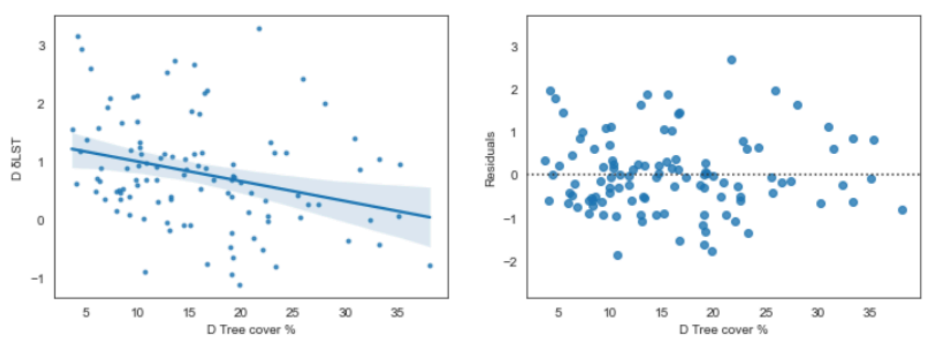
Out[9]:
```
count    106.000000
mean       0.789623
std        0.927848
min       -1.120000
25%        0.260000
50%        0.690000
75%        1.172500
max        3.280000
Name: D δLST, dtype: float64
```

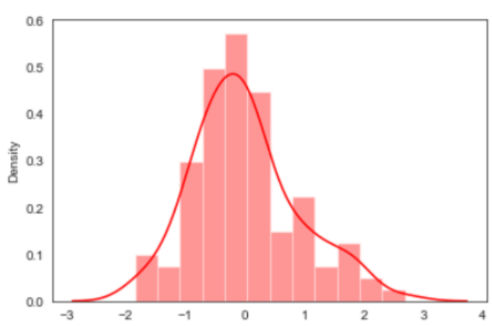
Appendix B - Data Visualizations (Scatterplots, Histogram, Correlation Matrix & Regression Parameters (R, R-squared, equations)

```
In [15]: data_1 = data.dropna()

         #THE BELOW IS TAKEN FROM THE CLASS SESSION 3.2 (SYNTHESIS)

         def regression_model(column_x, column_y):
             # this function uses built in library functions to create a scatter plot,
             # plots of the residuals, compute R-squared, and display the regression eqn

             # fit the regression line using "statsmodels" library:
             X = statsmodels.add_constant(data_1[column_x]) #why add_constant
             Y = data_1[column_y]
             regressionmodel = statsmodels.OLS(Y,X).fit() #OLS = "ordinary least squares"

             # extract regression parameters from model, rounded to 3 decimal places:
             Rsquared = round(regressionmodel.rsquared,3)
             slope = round(regressionmodel.params[1],3)
             intercept = round(regressionmodel.params[0],3)

             # make plots:
             fig, (ax1, ax2) = plt.subplots(ncols=2, sharex=True, figsize=(12,4))
             sns.regplot(x=column_x, y=column_y, data=data, marker=".", ax=ax1) # scatter plot
             sns.residplot(x=column_x, y=column_y, data=data, ax=ax2) # residual plot
             ax2.set(ylabel='Residuals')
             ax2.set_ylim(min(regressionmodel.resid)-1,max(regressionmodel.resid)+1)
             plt.figure(edgecolor='k') # histogram
             sns.distplot(regressionmodel.resid, kde=True, axlabel='Residuals', color='red')

             # print the results:
             print("R-squared = ",Rsquared)
             print("Regression equation: "+column_y+" = ",slope,"* "+column_x+" + ",intercept)
```

```
In [17]: regression_model ("D Tree cover %", "D δLST")
```

```
R-squared =  0.084
Regression equation: D δLST =  -0.034 * D Tree cover % +  1.32
```



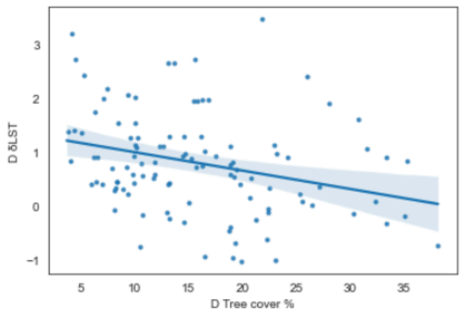


In [16]:
```
# calcualte r and make a scatter plot for two variables
def corr_scatter(column_a, column_b):
    print("\nThe pearson's r value comparing", column_a, "to", column_b, "is:",round(data_1[column_a].corr(data_1[column_b]),3))
    sns.regplot(x= column_a, y= column_b, data=data, marker=".", x_jitter=.25, y_jitter=.25)
    # jitter is added to offset data points that are potentially overlapping due to discreteness.

print("The corr_scatter(column_a,column_b) function is loaded.")
```

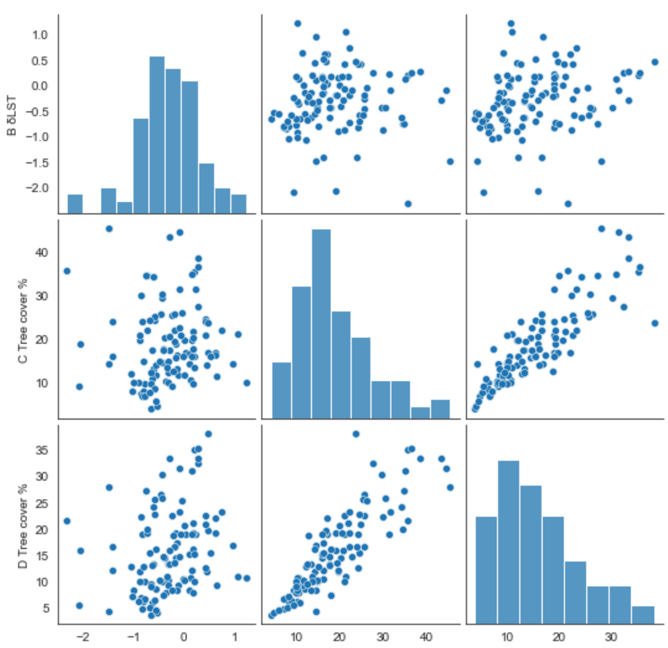
```
The corr_scatter(column_a,column_b) function is loaded.
```

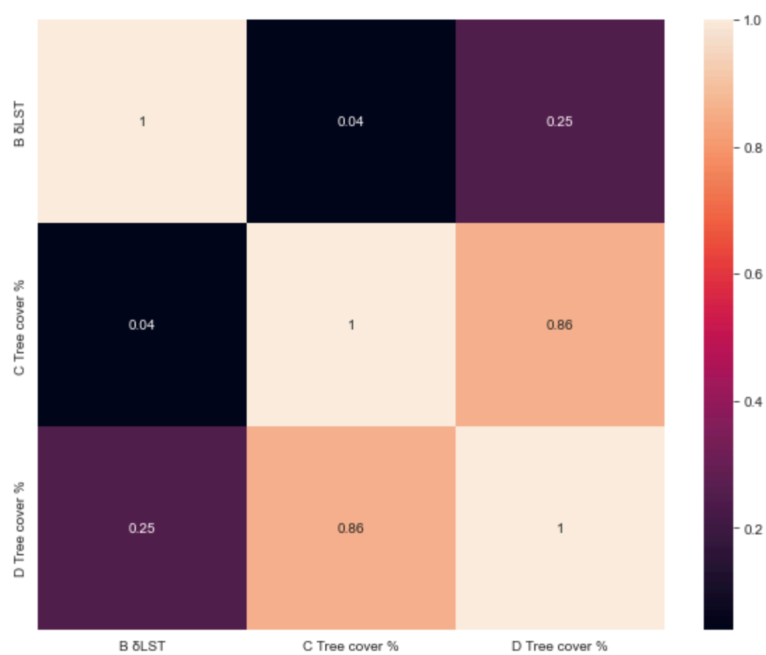
In [18]:
```
corr_scatter("D Tree cover %", "D δLST")
```

```
The pearson's r value comparing D Tree cover % to D δLST is: -0.291
```



In [37]:
```
#remove unrelevant variables, only check the correlation of B δLST, D Tree cover %, and C Tree cover %
data_2 = data.drop(columns = ['Lansat Date','Urban area', 'State', 'D-A (°C)', 'C δLST', 'D δLST','A δLST', 'A Tree cover %', 'B
corr_map_data_2 = data_2.corr().round(2) #find the pairwise correlation of all columns in the Pandas Dataframe in Python
corr_map_data_2

sns.pairplot(data_2) #create a correlation plot for a pairwise relationship of the variables
plt.figure(figsize=(10,8))
plot = sns.heatmap(data_2.corr().round(2), annot=True) #create a heatmap consisting of the corr.eff of each pair of variables
```





Appendix C - Significance Test (P-value & Confidence Interval)

In [16]:
```
from scipy import stats

r = -0.291 # correlation coefficient
x_bar = 15.845348 # mean of x-values
sx = 8.046114 # standard deviation of x-values
y_bar = 0.789623 # mean of y-values
sy = 0.927848 # standard deviation of y-values
n = 106 # sample size

b1 = (sy/sx)*r # slope of sample regression line
#standard deviation of y-values divide by the standard deviation of x-values, all of which multiply by corr.eff

print("b1 =",b1)

SE = (sy/sx) * ((1-r**2)/(n-2))**0.5 # standard error of the slope
#SD of y-values divide by SD of x-values, multiply by the squareroot of 1-R-squared (co.eff of determination)/(sample size-2)

print("SE =",SE)

t = (b1-0)/SE # t-statistic
T_corrected = abs(t) # take the absolute value of t
print("t =",t)

p = (1-stats.t.cdf(T_corrected,n-2))*2 # two-tailed test with degrees of freedom (df) = n-2
print("p =", round(p,4))
```

```
b1 = -0.03355703983314181
SE = 0.010818329964810646
t = -3.1018687673878103
p = 0.0025
```

In [17]:
```
from scipy import stats

# given summary statistics:
r = -0.291 # correlation coefficient
x_bar = 15.845348 # mean of x-values
sx = 8.046114 # standard deviation of x-values
y_bar = 0.789623 # mean of y-values
sy = 0.927848 # standard deviation of y-values
n = 106 # sample size

b1 = (sy/sx)*r # slope of sample regression line
#standard deviation of y-values divide by the standard deviation of x-values, all of which multiply by corr.eff

print("b1 =",b1)

SE = (sy/sx)*(((1-r**2)/(n-2))**(0.5)) # standard error of the slope
#SD of y-values divide by SD of x-values, multiply by the squareroot of 1-R-squared (co.eff of determination)/(sample size-2)

print("SE =",SE)

T = abs(stats.t.ppf(1-0.1/2,n-2)) # t critical value for two-tailed test with (1-alpha)/2 and degrees of freedom = n-2
print("t =",t)

lower_bound = b1 - T*SE
upper_bound = b1 + T*SE

print("interval =", [lower_bound,upper_bound])
```

```
b1 = -0.03355703983314181
SE = 0.010818329964810646
t = -3.1018687673878103
interval = [-0.051511545245457147, -0.015602534420826469]
```

Appendix D - Forward Selection (Adjusted R-squared)

In [38]:

```
# run this cell to define the variables and a useful function

# dependent variable:
Y = data_1['D δLST']

# subset of possible independent variables:
predictors_subset = ['A δLST','B δLST','C δLST','A Tree cover %','B Tree cover %','C Tree cover %', 'D Tree cover %']

# function to compute a list of adjusted R^2 values for each predictor
def Rsquared_finder(predictors_list):
    Rsquared_list = []
    for n in range(len(predictors_list[0])):
        if len(predictors_list)==1:
            X = data_1[predictors_list[0][n]]
        elif len(predictors_list)==2:
            X = data_1[[predictors_list[0][n],predictors_list[1]]]
        elif len(predictors_list)==3:
            X = data_1[[predictors_list[0][n],predictors_list[1],predictors_list[2]]]
        X = statsmodels.add_constant(X) # if excluded, the intercept would be 0
        model = statsmodels.OLS(Y, X).fit()
        Rsquared = model.rsquared_adj
        Rsquared_list.append(round(Rsquared,5))
    return Rsquared_list

print('Variables and functions are loaded')
```

```
Variables and functions are loaded
```

In [40]:

```
# call the R^2 function and output the results
from tabulate import tabulate
Rsquared_list = Rsquared_finder([predictors_subset])
headers = ['Predictor','Adj R^2']
print(tabulate(np.transpose([predictors_subset,Rsquared_list]),headers))
```

```
Predictor           Adj R^2
--------------  ---------
A δLST            0.07507
B δLST            0.27703
C δLST            0.00246
A Tree cover %    0.00504
B Tree cover %    0.00849
C Tree cover %   -0.00959
D Tree cover %    0.07561
```

In [14]:

```
predictor1 = "D Tree cover %" # insert predictor here as a string

X = data_1[predictor1]
X = statsmodels.add_constant(X) # if excluded, the intercept would be 0
model = statsmodels.OLS(Y, X).fit()
model.summary()
```

Dep. Variable:	D δLST	**R-squared:**	0.084
Model:	OLS	**Adj. R-squared:**	0.076
Method:	Least Squares	**F-statistic:**	9.588
Date:	Sun, 29 Jan 2023	**Prob (F-statistic):**	0.00252
Time:	17:13:44	**Log-Likelihood:**	-137.29
No. Observations:	106	**AIC:**	278.6
Df Residuals:	104	**BIC:**	283.9
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.3205	0.192	6.874	0.000	0.940	1.701
D Tree cover %	-0.0335	0.011	-3.096	0.003	-0.055	-0.012

In [20]:
```
predictor2 = "C Tree cover %" # insert new predictor here as a string
predictors_setof2 = [predictor1, predictor2]

X = data_1[predictors_setof2]
X = statsmodels.add_constant(X) # if excluded, the intercept would be 0
model = statsmodels.OLS(Y, X).fit()
model.summary()
```

Out[20]:

OLS Regression Results

Dep. Variable:	D δLST	**R-squared:**	0.312
Model:	OLS	**Adj. R-squared:**	0.299
Method:	Least Squares	**F-statistic:**	23.38
Date:	Sun, 29 Jan 2023	**Prob (F-statistic):**	4.25e-09
Time:	17:16:10	**Log-Likelihood:**	-122.13
No. Observations:	106	**AIC:**	250.3
Df Residuals:	103	**BIC:**	258.2
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.9906	0.177	5.610	0.000	0.640	1.341
D Tree cover %	-0.1258	0.018	-6.838	0.000	-0.162	-0.089
C Tree cover %	0.0951	0.016	5.841	0.000	0.063	0.127

In [31]:
```
predictor3 = "B δLST"
predictors_setof3 = [predictor1, predictor2, predictor3]

X = data_1[predictors_setof3]
X = statsmodels.add_constant(X) # if excluded, the intercept would be 0
model = statsmodels.OLS(Y, X).fit()
model.summary()
```

Out[31]:

OLS Regression Results

Dep. Variable:	D δLST	**R-squared:**	0.427
Model:	OLS	**Adj. R-squared:**	0.411
Method:	Least Squares	**F-statistic:**	25.38
Date:	Sun, 29 Jan 2023	**Prob (F-statistic):**	2.41e-12
Time:	17:42:42	**Log-Likelihood:**	-112.42
No. Observations:	106	**AIC:**	232.8
Df Residuals:	102	**BIC:**	243.5
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.7414	0.171	4.336	0.000	0.402	1.081
D Tree cover %	-0.0928	0.018	-5.049	0.000	-0.129	-0.056
C Tree cover %	0.0717	0.016	4.540	0.000	0.040	0.103
B δLST	-0.5514	0.122	-4.529	0.000	-0.793	-0.310

In [33]:
```
predictor4 = "A δLST"
predictors_setof4 = [predictor1, predictor2, predictor3, predictor4]

X = data_1[predictors_setof4]
X = statsmodels.add_constant(X) # if excluded, the intercept would be 0
model = statsmodels.OLS(Y, X).fit()
model.summary()
```

Out[33]:

OLS Regression Results

Dep. Variable:	D δLST	**R-squared:**	0.457
Model:	OLS	**Adj. R-squared:**	0.435
Method:	Least Squares	**F-statistic:**	21.21
Date:	Sun, 29 Jan 2023	**Prob (F-statistic):**	1.02e-12
Time:	17:44:36	**Log-Likelihood:**	-109.65
No. Observations:	106	**AIC:**	229.3
Df Residuals:	101	**BIC:**	242.6
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.5209	0.192	2.708	0.008	0.139	0.902
D Tree cover %	-0.0896	0.018	-4.963	0.000	-0.125	-0.054
C Tree cover %	0.0682	0.016	4.390	0.000	0.037	0.099
B δLST	-0.5140	0.120	-4.274	0.000	-0.753	-0.275
A δLST	-0.1312	0.056	-2.328	0.022	-0.243	-0.019