Statistical Inference

Minerva Schools at KGI

CS50: Formal Analyses Fall 2020

December 16, 2020

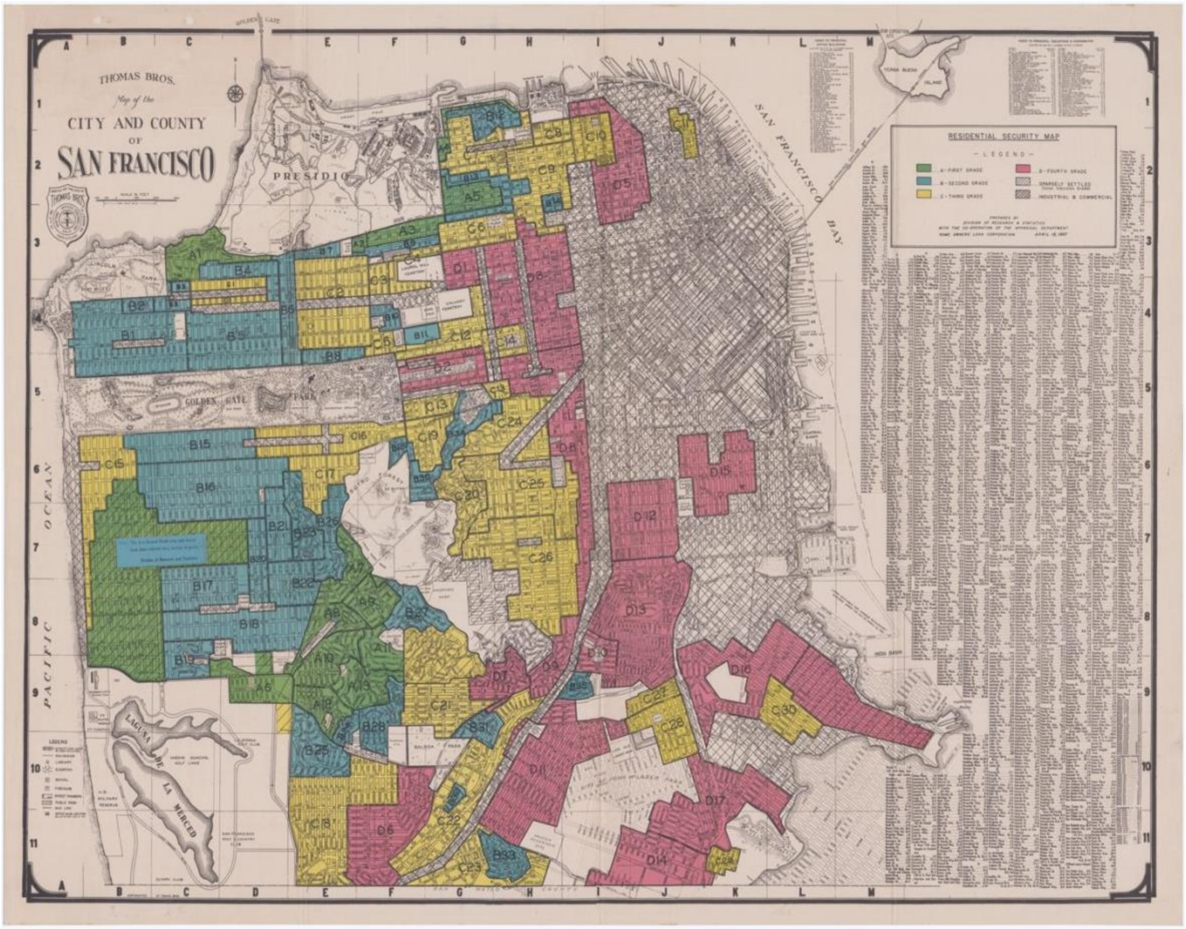
Statistical Inference

Part 1 Introduction

In the 1930s, to alleviate the negative economic effects of the Great Depression, a U.S. government program, Home Owners’ Loan Corporation (HOLC), created four categories of neighborhoods based primarily based on racial makeup: "Best" (category A), "Still Desirable" (category B), "Definitely Declining" (category C), and "Hazardous" (category D), as shown in Figure 1 (Hoffman et al., 2020). These distinctions led to systematic discrimination for people of color to make mortgage loans and secure homeownership.

Figure 1

Mapping Inequality: Redlining in New Deal America (San Francisco).



Note. The figure is retrieved from “Mapping Inequality,” American Panorama, and was produced by Robert K. Nelson, LaDale Winling, Richard Marciano, Nathan Connolly, et al. Different colors measure distinctions of neighborhoods in San Francisco, with green representing “Best” (category A), blue representing “Still Desirable” (category B), yellow representing “Definitely Declining” (category C), and red representing “Hazardous” (category D).

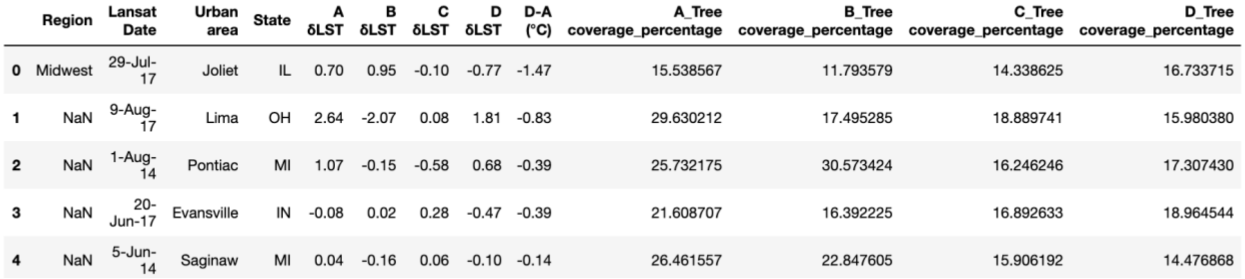
Other than housing impacts, redlining could also create temperature differences between different categories, as categories C and D tend to have less tree cover and more industrial plants (Hoffman et al., 2020). This report explores the temperature differences between category A and category D neighborhoods based on sample data from a scientific study. First, I estimate the average temperatures of category A and D neighborhoods by using the summary statistics and constructing a confidence interval. Then, I perform a significance test to determine whether category D neighborhoods have higher temperatures than category A neighborhoods (one-side test). The test result will be analyzed through both statistical significance and practical significance perspectives. As extra heat creates danger for weather-related diseases, this report will provide insights on environmental and healthcare policy decisions.

Part 2 Dataset

I obtained the dataset from the scientific study "*The effects of historical housing policies on resident exposure to intra-urban heat: A study of 108 US urban areas*" conducted by Hoffman et al. in 2020. As Figure 2 shows, the study records the land surface temperatures anomalies (δLST) and tree cover percentages of four categories in each urban area. δLST represents how much land surface temperature of areas in a given redlining category within a city deviates from the city's overall average land surface temperature. The deviations (anomalies) provide temperature differences between redlining categories comparable across cities. After clearing and processing the dataset, I use a sample of 106 urban areas in the report $(n = 106)$.

Figure 2

Redlining and Climate Change Dataset Preview.



	Region	Lansat Date	Urban area	State	A δLST	B δLST	C δLST	D δLST	D-A (°C)	A_Tree coverage_percentage	B_Tree coverage_percentage	C_Tree coverage_percentage	D_Tree coverage_percentage
0	Midwest	29-Jul-17	Joliet	IL	0.70	0.95	-0.10	-0.77	-1.47	15.538567	11.793579	14.338625	16.733715
1	NaN	9-Aug-17	Lima	OH	2.64	-2.07	0.08	1.81	-0.83	29.630212	17.495285	18.889741	15.980380
2	NaN	1-Aug-14	Pontiac	MI	1.07	-0.15	-0.58	0.68	-0.39	25.732175	30.573424	16.246246	17.307430
3	NaN	20-Jun-17	Evansville	IN	-0.08	0.02	0.28	-0.47	-0.39	21.608707	16.392225	16.892633	18.964544
4	NaN	5-Jun-14	Saginaw	MI	0.04	-0.16	0.06	-0.10	-0.14	26.461557	22.847605	15.906192	14.476868

Note. The figure only demonstrates the first five rows of the dataset. The codes of inputting and displaying the dataset will be provided in the appendix.

As the report compares temperature differences of category A and D neighborhoods, variables "A δLST" (land surface temperature anomalies of category A neighborhood) and "D δLST" (land surface temperature anomalies of category D neighborhood) are used. In nature, both variables are quantitative continuous. However, both variables are recorded as quantitative discrete with the smallest unit of 0.01 Celsius degree because of the artificial scale in temperature measurement. In the statistical context, the variables are the two samples I take.[1]

[1] **#variables**: I identified two key variables (A δLST and D δLST) for the following summary statistics, data visualizations, confidence interval, and significance test. I provided detailed descriptions of the variables (temperature anomalies) and their features (quantitative continuous in nature; quantitative discrete by measurements), and justified why I selected them (research question).

Part 3 Analysis

3.1 Summary Statistics and Data Visualization

To start the analysis, I first calculate the summary statistics of two samples (Table 1) and draw the sample distribution (Figure 3).

Table 1

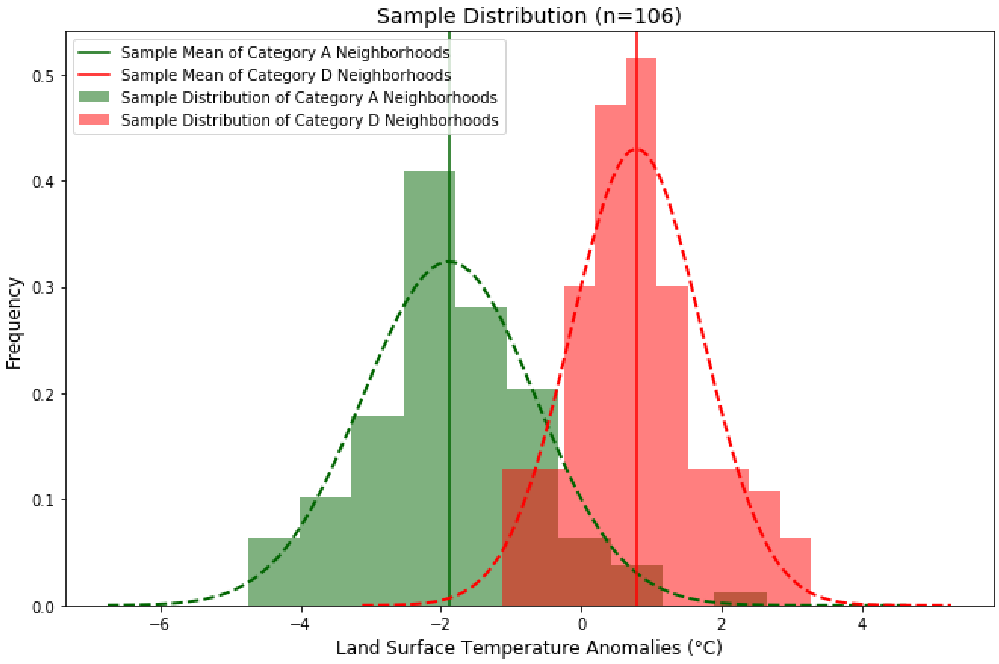
Summary Statistics of Two Samples

Summary Statistics	Category A Nneighborhoods sample	Category D Neighborhoods Sample
Count	106	106
Mean	-1.88	0.79
Standard Deviation	1.23	0.93
Median	-2.0	0.69
Range	7.39	4.4
Mode	-2.0	NA

Note. The Sample of Category D neighborhoods has no mode (two equally common values).

Figure 3

Sample Distribution



Based on the descriptive statistics, there is a relatively large difference between the sample mean of category A neighborhoods (-1.88) and that of category D neighborhoods (0.79), indicating a higher probability that category D neighborhoods have higher temperatures. Additionally, the sample of category A neighborhoods has a higher standard deviation (1.23 > 0.93) and range (7.39 > 4.4) than those of category D neighborhoods, which implies the temperatures of category A neighborhoods have higher variability. [2]

Each sample distribution is unimodal with small skewness and positive kurtosis compared to the normal distribution curve (dashed curve) of the same sample mean and standard deviation from the sample distribution figure. Furthermore, the difference between sample means can be visualized through the two vertical lines (the means of two samples). However, a formal, rigorous significance test is still needed for its statistical and practical significance value (which will be presented later). [3][4]

[2] **#descriptivestats**: I chose appropriate summary statistics (mean, standard deviation, median, mode, range) and justified the choice through interpretation of each #descriptivestats. I calculated the statistics using Python (Appendix B) and provided detailed interpretation of them (higher mean – the alternative hypothesis is more likely to be true; higher standard deviation and larger range – higher variability).

[3] **#dataviz**: I used python matplotlib package (Appendix B) to generate an appropriate histogram showing the sample distributions. I labeled the lines, distributions, and axes. Additionally, I adjusted the opacity value to see the overlapping of two sample distributions and added normal distribution curves, vertical lines of two sample means.

[4] **#distributions**: I plotted the two sample distributions using histogram. Additionally, I adjusted the opacity value to see the overlapping of two sample distributions and added normal distribution curves, vertical lines of two sample means. I described the characteristics of the distributions (unimodal; small skewness; positive kurtosis) and analyzed how the distribution’s features inform us about the significance test.

3.2 Confidence Interval

Using the samples to estimate the average temperatures of category A and D neighborhoods, I compute two confidence intervals that provide plausible ranges of values. The confidence level is set at the standard value of 95%. I use the T-statistics here because I do not know the real population standard deviation. I estimate the standard error using the sample standard deviation. T-distributions' tails are thicker than the normal distributions', and they are precisely the correction needed to resolve the problem of a poorly estimated standard error.

Three conditions need to be met before using the T-statistics. First, randomness. The dataset description in the original scientific study does not specify whether the researchers took a random sample. However, the datasets include neighborhoods of North, South, West, and East evenly. Therefore, there is a relatively high probability that this is a random sample. Second, the sampling distribution of the sample mean should be normal. This condition holds because the sample size (106) is significantly bigger than 30. Third, individual observations need to be independent. In the sampling without replacement case, the sample size (106) is less than 10% of the population size (3573) (U.S. Census Bureau, 2010). Thus, all three conditions for inference on a mean are met.

Table 2

Confidence Interval Calculation

	Formula	**Category A Neighborhoods Sample**	**Category D Neighborhoods Sample**
Standard Error	$\frac{\sigma_{sample}}{\sqrt{n}}$	0.120	0.090
Margin of Error	$\frac{\sigma_{sample}}{\sqrt{n}} \times t$	0.237	0.179
Confidence Interval	$mean \pm \frac{\sigma_{sample}}{\sqrt{n}} \times t$	(-2.120, -1.645)	(0.611, 0.968)

Confidence Interval after the Finite Population Correction	$mean \pm \frac{\sigma_{sample}}{\sqrt{n}} \times t \times \sqrt{\frac{N-n}{N-1}}$	(-2.116, -1.649)	(0.614, 0.966)

Table 2 shows the confidence intervals of two samples with and without the Finite Population Correction (FPC). FPC is needed because the standard error formula relies on the assumption that there is an infinite population. However, the population size is finite (3573). As the FPC $\sqrt{\frac{N-n}{N-1}}$ is always smaller than 1, it would effectively decrease the standard error, increasing accuracy when sampling without replacement. In the context, the calculation signifies that we are 95% confident that the range between -2.116 and -1.649 captures the average temperature of Category A neighborhoods, and the range between 0.614 and 0.966 captures the average temperature of Category D neighborhoods. [5]

[5] **#confidenceintervals**: I calculated two confidence intervals for two given samples and population characteristic with clear and detailed steps (Appendix C). I provided interpretation of the two confidence intervals (95% that the range captures the real population mean). Additionally, I applied FPC (finite population correction) to reach more accurate confidence interval values

3.3 Difference of Means Significance Test

To address my research question of whether Category D neighborhoods have higher temperatures than Category A neighborhoods, I conduct a difference of means significance test using two samples. The significance level α is set to the standard value of 0.05. The test has the following hypotheses:

Null Hypothesis

Category A neighborhoods and Category D neighborhoods have the same average temperatures.

$$\mu_A - \mu_D = 0$$

Alternative Hypothesis

Category D neighborhoods have a higher average temperature than that of Category A neighborhoods.

$$\mu_D - \mu_A > 0$$

Based on the hypotheses, the test is one-tailed, as I only look for the direction where Category D neighborhoods' average temperature is greater than that of A. The three conditions are met (checked in 3.2). Similarly, because the population standard deviations are unknown, I use the T-statistics.

$$df\ (degrees\ of\ freedom) = min(n_A - 1, n_D - 1)$$

$$SE_{(\bar{x}_A - \bar{x}_D)} = \sqrt{\frac{s_A^2}{n_A} + \frac{s_D^2}{n_D}}$$

$$point\ estimate \pm margin\ of\ error: (\bar{x}_A - \bar{x}_D) \pm t_{df} \times SE_{(\bar{x}_A - \bar{x}_D)}$$

The calculation shows that $p = 7.59 \times 10^{-34}$. The value is significantly smaller than 0.05. Therefore, I reject the null hypothesis and take the alternative hypothesis. The effect is statistically significant because it exists. Then, I use Cohen's d effective size to evaluate its practical significance. I choose Cohen's d because first, it offers a scale to evaluate the effect size as shown in Table 3, and second, alternative measure Glass’s Delta is more suitable for clinical studies where there is a control group and an interventional group.

$$Cohen's\ d = \frac{\overline{x}_A - \overline{x}_D}{\sqrt{\frac{(n_A - 1)s_A^2 + (n_D - 1)s_D^2}{n_A + n_D - 2}}}$$

Table 3

Descriptors for Magnitudes of Cohen’s d value

Cohen’s d value	**Effective Size**
0.01	Very small
0.20	Small
0.50	Medium
0.80	Large
1.20	Very large
2.0	Huge

Note. The table was initially suggested by Cohen and later expanded by Sawilowsky in “*New effect size rules of thumb*," Journal of Modern Applied Statistical Methods (2009).

The calculation shows that $d = 2.4$. This is a huge effect size based on Table 3. It indicates that there is a huge practical significance. In the context, this signifies that the average temperature of Category D neighborhoods is remarkedly larger than that of Category A neighborhoods.[6]

[6] **#significance**: I clearly stated and explained the null hypothesis and alternative hypothesis (one-side). I conducted the significance test with detailed, justified calculation (Appendix D) and provided the formula. I calculated and evaluated both statistical significance (p-value < significance level α) and practical significance (Cohen’s d effect size) values and interpreted the results in the redlining context.

Part 4 Result and Conclusions

To summarize, based on $p = 7.59 \times 10^{-34}$ and $d = 2.4$, the null hypothesis should be rejected (statistically significant). As Cohen's d effect size is huge, the result is also practically significant. The average temperature of Category D neighborhoods (classified as "Hazardous") is inducted to be notably higher than the average temperature of Category A neighborhoods (classified as "Best"). The result highlights the government's need to devote more resources to environmental protection and climate change for Category D neighborhoods. Also, the result serves as a realistic reflection of systematic discrimination in the U.S. The negative impacts brought by discrimination is far more pervasive than we could imagine.

The inference in this report is inductive because the conclusions are made based on two samples. A generalization induction is used when we infer the population's characteristics based on the samples. The induction is relatively strong because all three conditions (assumptions) for using T-statistics and significance tests are met. Second, there is a use of multiple corrections to enhance the results' accuracy. However, the reliability of the induction could be questioned. Further information about whether the two samples represent the whole population is unknown (some premises' quality is unknown). Future studies could employ more samples to reach higher accuracy.[7]

[7] **#induction**: I applied induction to draw conclusions/inference on the population characteristics and their significance in real world (1. Environmental and health concerns 2. Reflection of systematic discrimination). I justified the process of applying induction and the type of induction use (generalization; using the sample characteristics to infer the population characteristics). I evaluated the strength and reliability of the induction and offered future direction to improve the induction’s strength/reliability.

Part 5 Reflection

My TA Ha Tran Nguyen Phuong's reflection poll feedback of Session 24 Effective Size and Difference of Mean Tests is valuable. Ha Tran suggests that the assignment must include BOTH a measure of statistical significance (p-value) and practical significance (effect size, like d or g). Not only do we want to know whether we have sufficient statistical evidence to reject the null (p-value < alpha), we would also like to see the size of that effect. Therefore, in my assignment, I included both the p-value for statistical significance and Cohen's d value for practical significance. The two measures complement each other and offer me thorough insights tailored to the context of redlining. This experience is reflected in the #scienceoflearning principle "deliberate practice," where we learn through feedback and correct our mistakes along the way.

Word Count: 149 words (excluding Part 5 Reflection)[8]

[8] **#professionalism**: I followed established guidelines to present my work professionally: format (cover page, align left, bolding, title, font size, footnotes, page breaks, word count), grammar (checked by Grammarly in academic standards), tone (academic language), reference (in APA citation format), and appendix (all the python codes used). All the tables and figures are labeled with numbers, titles, and (if applicable) brief notes/captions.

References

Hoffman, J. S., Shandas, V., and 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, 12. doi:10.3390/cli8010012. https://www.mdpi.com/2225-1154/8/1/12

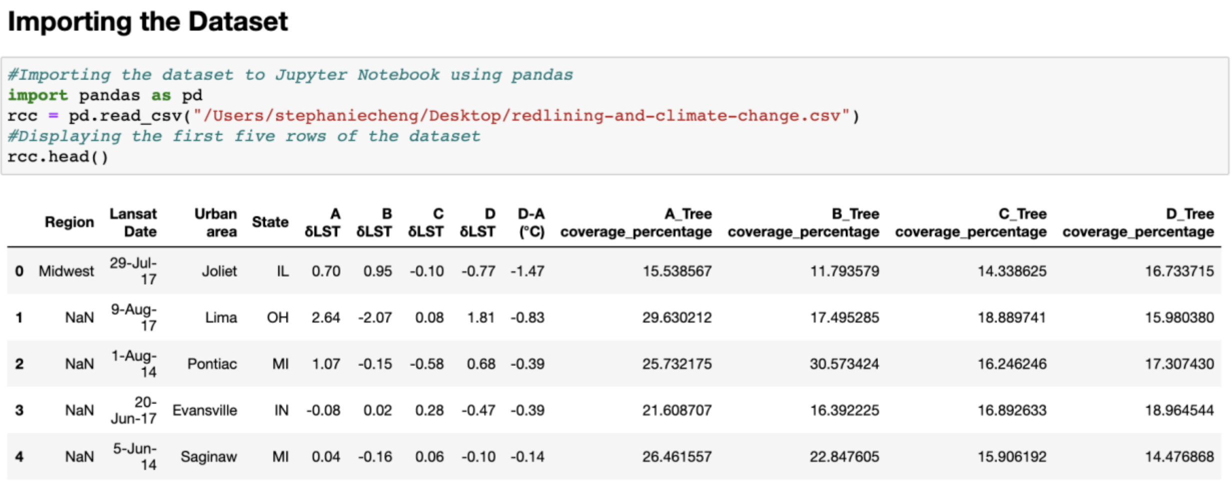
U.S. Census Bureau. (2010). *Urban, Urbanized Area, Urban Cluster, and Rural Population, 2010 and 2000: United States*. Urban Area Facts. https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural/ua-facts.html

Sawilowsky, S. (2009). “*New effect size rules of thumb*." Journal of Modern Applied Statistical Methods. 8 (2): 467–474. doi:10.22237/jmasm/1257035100. http://digitalcommons.wayne.edu/jmasm/vol8/iss2/26/

Appendix

Appendix A

Part 2 Dataset



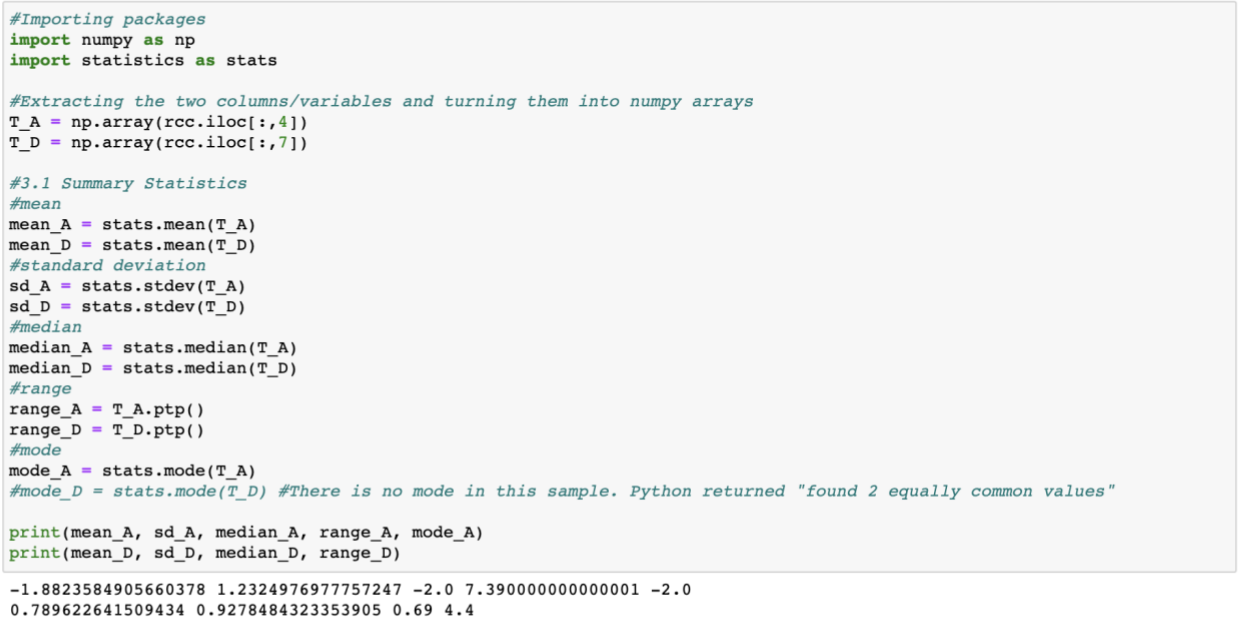
Importing the Dataset

```
#Importing the dataset to Jupyter Notebook using pandas
import pandas as pd
rcc = pd.read_csv("/Users/stephaniecheng/Desktop/redlining-and-climate-change.csv")
#Displaying the first five rows of the dataset
rcc.head()
```

	Region	Lansat Date	Urban area	State	A δLST	B δLST	C δLST	D δLST	D-A (°C)	A_Tree coverage_percentage	B_Tree coverage_percentage	C_Tree coverage_percentage	D_Tree coverage_percentage
0	Midwest	29-Jul-17	Joliet	IL	0.70	0.95	-0.10	-0.77	-1.47	15.538567	11.793579	14.338625	16.733715
1	NaN	9-Aug-17	Lima	OH	2.64	-2.07	0.08	1.81	-0.83	29.630212	17.495285	18.889741	15.980380
2	NaN	1-Aug-14	Pontiac	MI	1.07	-0.15	-0.58	0.68	-0.39	25.732175	30.573424	16.246246	17.307430
3	NaN	20-Jun-17	Evansville	IN	-0.08	0.02	0.28	-0.47	-0.39	21.608707	16.392225	16.892633	18.964544
4	NaN	5-Jun-14	Saginaw	MI	0.04	-0.16	0.06	-0.10	-0.14	26.461557	22.847605	15.906192	14.476868

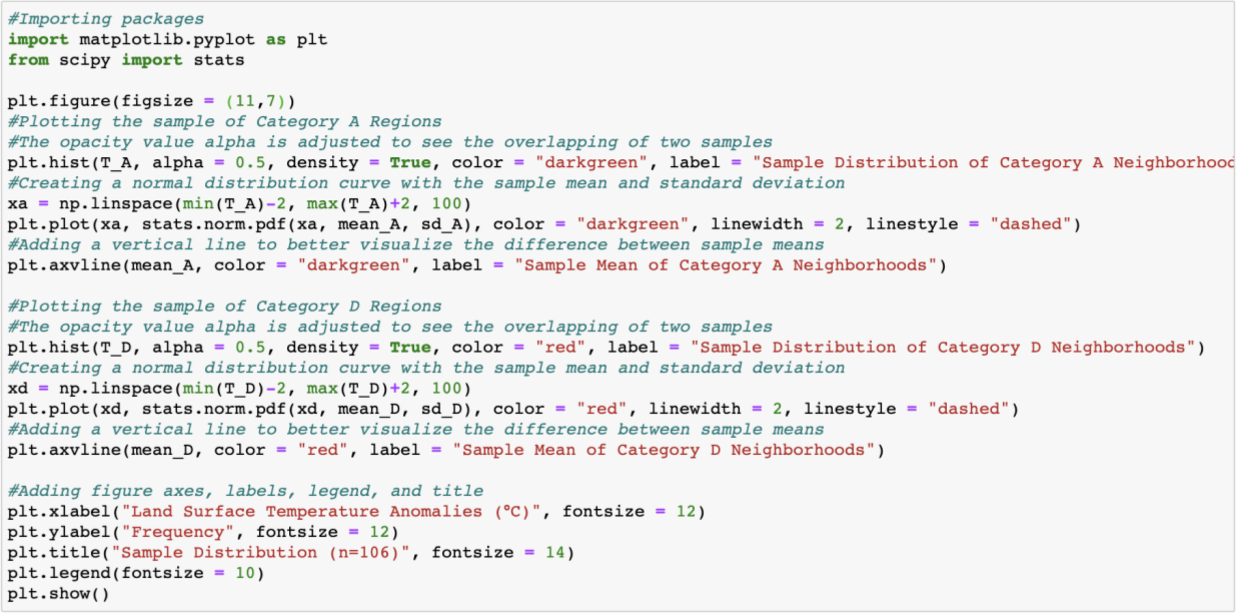
Appendix B

Part 3 – 3.1 Summary Statistics and Data Visualization



```
#Importing packages
import numpy as np
import statistics as stats

#Extracting the two columns/variables and turning them into numpy arrays
T_A = np.array(rcc.iloc[:,4])
T_D = np.array(rcc.iloc[:,7])

#3.1 Summary Statistics
#mean
mean_A = stats.mean(T_A)
mean_D = stats.mean(T_D)
#standard deviation
sd_A = stats.stdev(T_A)
sd_D = stats.stdev(T_D)
#median
median_A = stats.median(T_A)
median_D = stats.median(T_D)
#range
range_A = T_A.ptp()
range_D = T_D.ptp()
#mode
mode_A = stats.mode(T_A)
#mode_D = stats.mode(T_D) #There is no mode in this sample. Python returned "found 2 equally common values"

print(mean_A, sd_A, median_A, range_A, mode_A)
print(mean_D, sd_D, median_D, range_D)
```

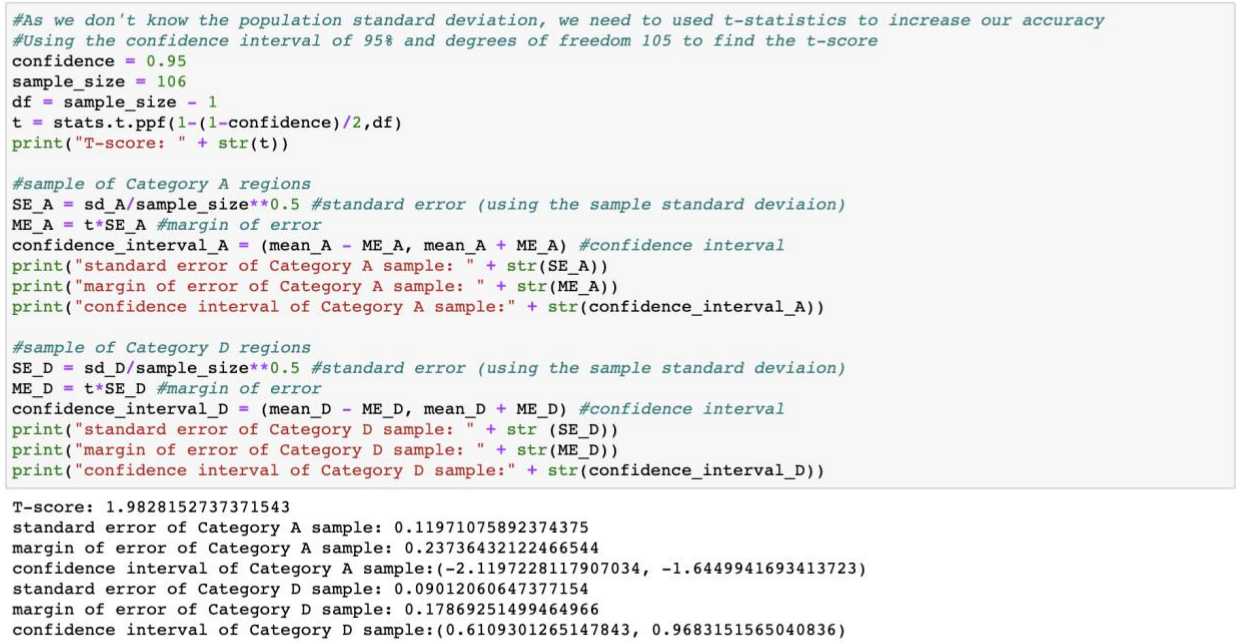
```
-1.8823584905660378 1.2324976977757247 -2.0 7.390000000000001 -2.0
0.789622641509434 0.9278484323353905 0.69 4.4
```



```
#Importing packages
import matplotlib.pyplot as plt
from scipy import stats

plt.figure(figsize = (11,7))
#Plotting the sample of Category A Regions
#The opacity value alpha is adjusted to see the overlapping of two samples
plt.hist(T_A, alpha = 0.5, density = True, color = "darkgreen", label = "Sample Distribution of Category A Neighborhoo
#Creating a normal distribution curve with the sample mean and standard deviation
xa = np.linspace(min(T_A)-2, max(T_A)+2, 100)
plt.plot(xa, stats.norm.pdf(xa, mean_A, sd_A), color = "darkgreen", linewidth = 2, linestyle = "dashed")
#Adding a vertical line to better visualize the difference between sample means
plt.axvline(mean_A, color = "darkgreen", label = "Sample Mean of Category A Neighborhoods")

#Plotting the sample of Category D Regions
#The opacity value alpha is adjusted to see the overlapping of two samples
plt.hist(T_D, alpha = 0.5, density = True, color = "red", label = "Sample Distribution of Category D Neighborhoods")
#Creating a normal distribution curve with the sample mean and standard deviation
xd = np.linspace(min(T_D)-2, max(T_D)+2, 100)
plt.plot(xd, stats.norm.pdf(xd, mean_D, sd_D), color = "red", linewidth = 2, linestyle = "dashed")
#Adding a vertical line to better visualize the difference between sample means
plt.axvline(mean_D, color = "red", label = "Sample Mean of Category D Neighborhoods")

#Adding figure axes, labels, legend, and title
plt.xlabel("Land Surface Temperature Anomalies (°C)", fontsize = 12)
plt.ylabel("Frequency", fontsize = 12)
plt.title("Sample Distribution (n=106)", fontsize = 14)
plt.legend(fontsize = 10)
plt.show()
```

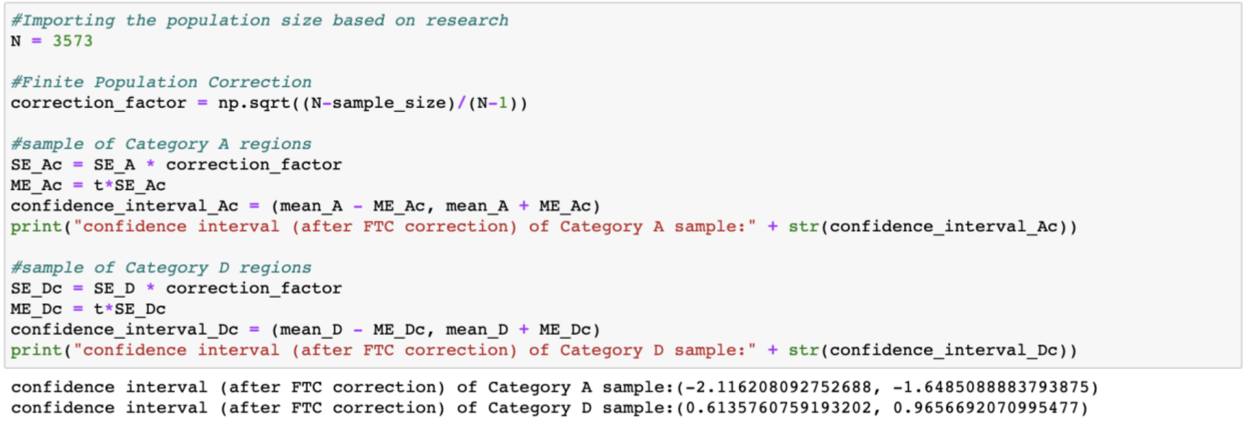
Appendix C

Part 3 – 3.2 Confidence Interval



```
#As we don't know the population standard deviation, we need to used t-statistics to increase our accuracy
#Using the confidence interval of 95% and degrees of freedom 105 to find the t-score
confidence = 0.95
sample_size = 106
df = sample_size - 1
t = stats.t.ppf(1-(1-confidence)/2,df)
print("T-score: " + str(t))

#sample of Category A regions
SE_A = sd_A/sample_size**0.5 #standard error (using the sample standard deviaion)
ME_A = t*SE_A #margin of error
confidence_interval_A = (mean_A - ME_A, mean_A + ME_A) #confidence interval
print("standard error of Category A sample: " + str(SE_A))
print("margin of error of Category A sample: " + str(ME_A))
print("confidence interval of Category A sample:" + str(confidence_interval_A))

#sample of Category D regions
SE_D = sd_D/sample_size**0.5 #standard error (using the sample standard deviaion)
ME_D = t*SE_D #margin of error
confidence_interval_D = (mean_D - ME_D, mean_D + ME_D) #confidence interval
print("standard error of Category D sample: " + str (SE_D))
print("margin of error of Category D sample: " + str(ME_D))
print("confidence interval of Category D sample:" + str(confidence_interval_D))
```

```
T-score: 1.9828152737371543
standard error of Category A sample: 0.11971075892374375
margin of error of Category A sample: 0.23736432122466544
confidence interval of Category A sample:(-2.1197228117907034, -1.6449941693413723)
standard error of Category D sample: 0.09012060647377154
margin of error of Category D sample: 0.17869251499464966
confidence interval of Category D sample:(0.6109301265147843, 0.9683151565040836)
```

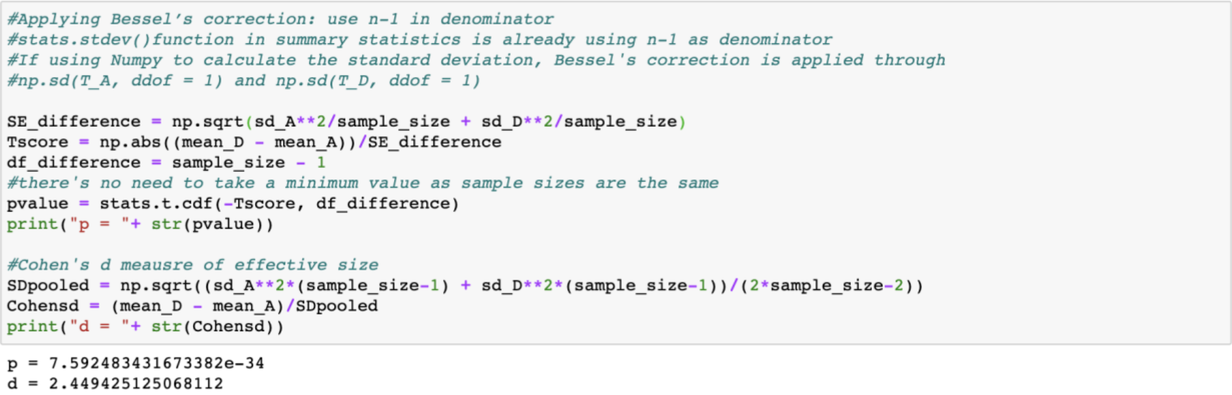


```
#Importing the population size based on research
N = 3573

#Finite Population Correction
correction_factor = np.sqrt((N-sample_size)/(N-1))

#sample of Category A regions
SE_Ac = SE_A * correction_factor
ME_Ac = t*SE_Ac
confidence_interval_Ac = (mean_A - ME_Ac, mean_A + ME_Ac)
print("confidence interval (after FTC correction) of Category A sample:" + str(confidence_interval_Ac))

#sample of Category D regions
SE_Dc = SE_D * correction_factor
ME_Dc = t*SE_Dc
confidence_interval_Dc = (mean_D - ME_Dc, mean_D + ME_Dc)
print("confidence interval (after FTC correction) of Category D sample:" + str(confidence_interval_Dc))
```

```
confidence interval (after FTC correction) of Category A sample:(-2.116208092752688, -1.6485088883793875)
confidence interval (after FTC correction) of Category D sample:(0.6135760759193202, 0.9656692070995477)
```

Appendix D

Part 3 – 3.3 Significance Test



```
#Applying Bessel’s correction: use n-1 in denominator
#stats.stdev()function in summary statistics is already using n-1 as denominator
#If using Numpy to calculate the standard deviation, Bessel's correction is applied through
#np.sd(T_A, ddof = 1) and np.sd(T_D, ddof = 1)

SE_difference = np.sqrt(sd_A**2/sample_size + sd_D**2/sample_size)
Tscore = np.abs((mean_D - mean_A))/SE_difference
df_difference = sample_size - 1
#there's no need to take a minimum value as sample sizes are the same
pvalue = stats.t.cdf(-Tscore, df_difference)
print("p = "+ str(pvalue))

#Cohen's d meausre of effective size
SDpooled = np.sqrt((sd_A**2*(sample_size-1) + sd_D**2*(sample_size-1))/(2*sample_size-2))
Cohensd = (mean_D - mean_A)/SDpooled
print("d = "+ str(Cohensd))
```

```
p = 7.592483431673382e-34
d = 2.449425125068112
```