Data 102 Final Project

Group Members: Sohail Attari, Ankita Janakiraman, Kashish Kharbanda, Ishika Prashar

Fall 2022

Table of Contents

Data Overview 3

Research Questions 4

Research Question 1: Are racist redlining practices correlated with ozone and PM2.5 concentration levels? 4

Research Question 2: Can PM2.5 and ozone concentration levels predict racist neighborhood homeowner loan corporation grades? 4

EDA 4

Data Cleaning and Impact 4

Multiple hypothesis testing/decision making 7

Methods 7

Results 8

Discussion 11

Prediction with GLMs and Nonparametric methods 12

Methods 12

Results 12

Discussion 19

Conclusion 19

References 21

Data Overview

We used “Dataset 1: Chronic Disease and Air Quality.” Specifically, we used *CDC Daily Census - Tract PM2.5 Concentrations* and *CDC Daily Census - Tract Ozone Concentrations*. Both of these datasets generated PM2.5 and ozone levels from the EPA Downscaler model from 2011 to 2014. The Downscaler model combines output from other measurements and a separate air quality model. For the external dataset that we incorporated into our project, we downloaded our dataset from data.diversitydatakids.org (Home Owner Loan Corporation (HOLC) neighborhood grades for US census tracts). This is an organization that strives for equity for kids and has a plethora of relevant data. The chosen dataset was in the form of a csv, and an account was needed in order to download the file from the website.

We chose to use additional data because it provides different grade levels (A- best, B- still desirable, C- definitely declining, D- hazardous) assigned to neighborhoods for credit risk. This grading system was biased towards minority neighborhoods, so we will use this for our questions about possible correlations between racist redlining practice along with pollutant concentrations in US neighborhoods. The HOLC neighborhood grades data was collected in the 1930s. Hence, there could have been excluded groups who lived in non-traditional housing conditions such as homeless camps, etc who were not approached.

The 1930s HOLC neighborhood grades are mapped to more recent census tracts from 2010 and 2020. Redlining continues to have many repercussions today. The sources do not explicitly mention whether or not participants/neighborhoods were aware of the grades they were receiving. For census information, usually representatives from the Census Bureau will contact individuals. However, the sources do not explicitly mention whether or not details regarding the data use were shared with the neighborhoods.The granularity of the data is by ‘ctfipscode’ for a particular year. The ‘ctfipscode’ encompasses the state, county within the state, and the tract code combined. Since three datasets were merged, namely the PM2.5 concentration dataset, the ozone concentration dataset, and the redlining dataset, each row contains information for a particular ‘ctfipscode’ location regarding all of these attributes. This impacts the interpretation of the findings because then location based information for specific neighborhoods can be used for analysis.

All of the datasets were gathered from census data, and a variety of neighborhoods are included in this dataset. Although not explicitly mentioned, it is possible for selection bias to arise if neighborhoods and locations are not sampled appropriately. Measurement errors can arise when recording the ozone and PM2.5 concentrations in the neighborhoods. The PM2.5 and ozone concentrations were measured using the CDC National Environmental Public Health Network. Hence, if measurement errors are present, they will arise from inadequate use of this network. However, standard error is included in the datasets for every measurement that is collected.

Although we have all of the columns we need for modeling and answering our questions, other important features/columns that might have been helpful to strengthen our analysis include more supplemental information about the environmental conditions regarding pollution and air quality in these neighborhoods. For example, if there were measurements about more specific pollutant concentrations, this information could support and potentially correlate with the PM2.5 and ozone concentration information we already have.

Research Questions

Research Question 1: Are racist redlining practices correlated with ozone and PM2.5 concentration levels?

If we find that redlining practices are correlated with ozone and PM2.5 concentration levels, investigations by environmental groups can be conducted to determine why this is the case. Furthermore, during redlining, certain services are not provided to neighborhoods because they are considered hazardous, so the government can intervene to find out which organizations are responsible for increasing the concentration levels of ozone and PM2.5. These organizations can be told to reduce their pollution or face heavy financial penalties. The method we utilized is multiple hypothesis testing. This is a fit because we have multiple variables including grades for neighborhoods and concentration levels. By conducting multiple hypothesis tests, we can make accurate data-driven decisions. If our p-value is below the threshold we set, we reject the null hypothesis and determine that there is a difference in ozone or PM2.5 concentration levels between neighborhood grades. Hence, there is evidence to enact new policies or financial penalties.

Research Question 2: Can PM2.5 and ozone concentration levels predict racist neighborhood homeowner loan corporation grades?

Answering this question can help lead to changes for future neighborhood lining practices. If PM2.5 and ozone concentrations are accurate predictors of neighborhood Homeowner Loan Corporation grades, this would show that racist redlining practices influence minority neighborhoods to have to deal with environmental air quality issues. This predictor can be used to find such instances, and change future policies for loan corporations' grades. If we find that air quality is indeed worse in neighborhoods with poor grades, further action can be taken to alleviate the stressors of this situation for residents of these neighborhoods. Using GLMs and nonparametric methods is a good fit for this question because we are trying to predict grade levels based on numerical features. The ideal way to make predictions would be through generalized linear models or nonparametric methods such as random forests because we can use the provided data to train models and use this data to make predictions for a given set of concentration levels.

EDA

Data Cleaning and Impact

Three different datasets are used in this project; namely, the PM2.5 concentrations, the ozone concentrations, and the redlining dataset. Initially upon looking through the data and the relevant features, we found that all three datasets have a matching feature, Census tract FIPS code. In order to merge the datasets, we grouped the PM2.5 and ozone datasets by these FIPS codes Then, we averaged the concentration levels and their standard errors . Next, we merged all three datasets on the FIPS code. No other cleaning steps were necessary. Due to aggregating the concentration data by the census tract FIPS code, we lost the temporal data of the concentrations. This impacts our model and inferences because we are unable to model temporal changes, which may lead to a loss in additional insight in changes of concentration levels and regulations over time.

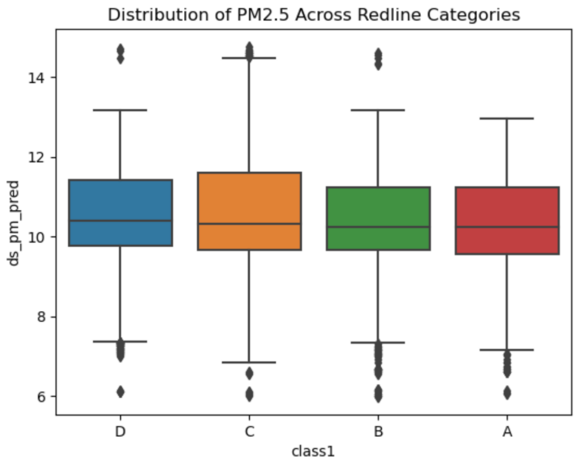


Figure 1a: Distribution of PM2.5 Across Redline Categories

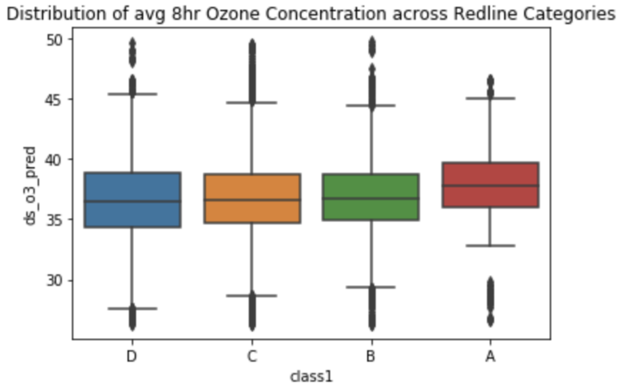


Figure 1b: Distribution of e Ozone Concentration across redline categories

Figure 1a/b:

Both box and whisker plot distributions underestimate the expected trend in their distributions of concentration levels. For PM2.5, we expected there to be higher pollutant concentrations in areas graded as ‘D’ or ‘C,’ and although there is a slight increase overall, especially in the ‘C’ grade areas, it is not substantial. Similarly, for ozone concentrations, the concentrations are uniform. Contrary to what we expected, the ozone concentrations are slightly higher at area ‘A.’ The data becomes slightly more spread out as we go from class ‘A’ to ‘D’ with the outliers in the data increasing, but average ozone concentrations tend to stay about the same amongst classes ‘B’, ‘C’, and ‘D.’

The boxplot demonstrates that through an initial look, there doesn’t seem to be a very strong correlation between redlining practices that were racially biased in creation and the Ozone/PM 2.5 concentrations. However, there is very clear evidence in Figure 1a that for the areas graded as ‘C’, the pollutant concentrations are higher. Additionally, it is possible that the outliers present in the pollutant concentrations may help predict racist redlining areas.

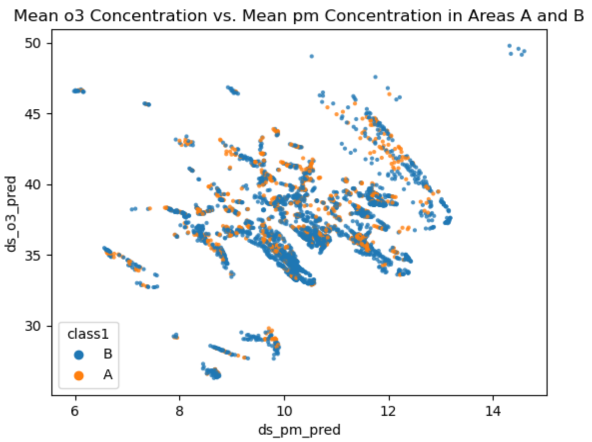


Figure 2a: Comparison of Ozone vs. PM2.5 Across Redline Grades A, B

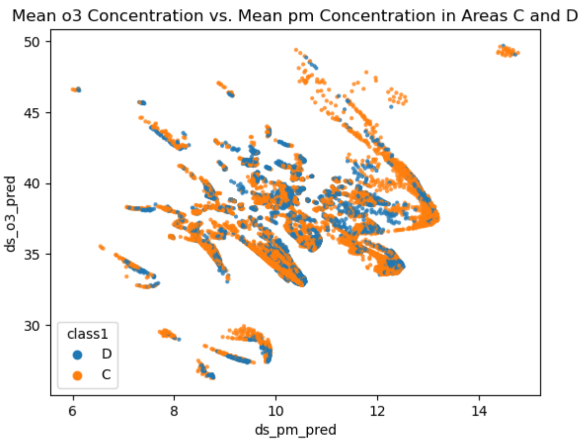
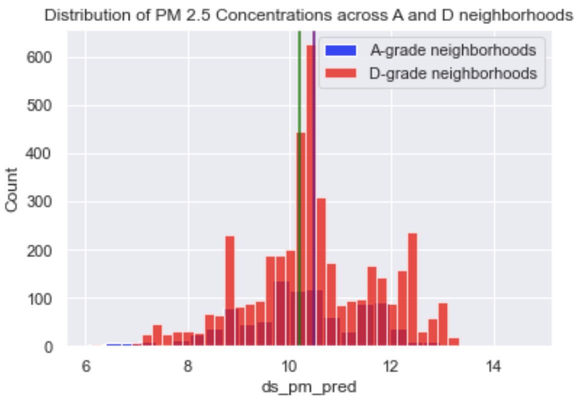


Figure 2b: Comparison of Ozone vs. PM2.5 Across Redline Grades C, D

Figure 2a/2b:

For all redline categories A, B, C, and D there appears to be a slight positive correlation between the PM2.5 and ozone concentrations. The scatterplot across areas with higher grades (A and B) appears to look similar to the scatterplot across areas with lower grades (C and D). We will follow up on distinguishing between the concentration levels and grades by accounting for location data and looking into areas with different grades with similar location (latitude, longitude, or fips code) attributes.

The scatter plots show the correlation between the ozone concentration and the PM2.5 concentration across both the high and low grade regions. These suggest potential directions for the research questions posed because the scatter plots appear to have the same trends and axis scales for concentrations in both the high and low grade neighborhoods. This suggests that beyond concentrations, there is something else contributing to these scores. For the second research question, these plots suggest that these concentration levels alone cannot predict the grades, and other information must be analyzed to understand score assignments.



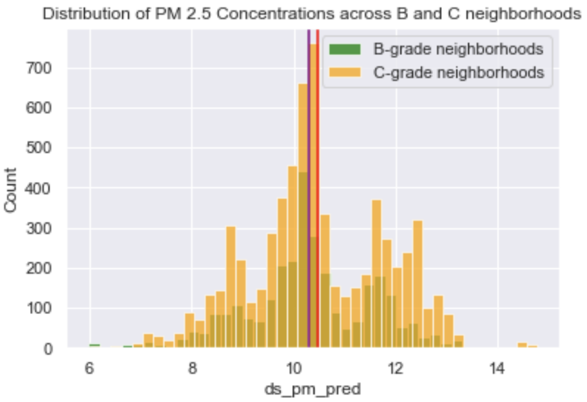


Figure 3a: Distribution PM 2.5 Across A & D Neighborhoods green line = mean of A-grade ; purple line = mean of D grade

Figure 3b: Distribution PM 2.5 Across B & C Neighborhoods purple line = mean of B-grade ; red line = mean of C-grade

Figure 3a/3b:

Due to overplotting, there are two histograms where A and D neighborhoods are overlaid and B and C neighborhoods are overlaid. In Figure 3A, we compare the extremes of the neighborhoods, expecting that 'A' grade areas that are known to be best would have a distribution of PM2.5 concentrations shifted more to the left than 'D' grade areas that are known to be hazardous. However, these histograms are very closely overlaid on top of each other, and we don't see a significant difference in the distribution of 'D' grade PM2.5 concentration levels shifted to the right. Furthermore, the vertical lines on the histogram depict the means of each of the grades and are almost identical. In Figure 3B, we expected there to be a smaller difference than between the A and D neighborhoods for the distribution of PM2.5 levels, but we can observe that the overlaid histograms for B and C neighborhoods are also almost identical. One additional technique to use is geographical heat maps based on the longitude and latitude variables. We can also explore the variables for percentage of non-water areas based on each neighborhood class and establish a correlation between that and the PM2.5 concentration levels.

Our first research question describes if we can establish a correlation between racist redlining practices and PM 2.5 concentration levels. Based on the 2 histograms, there doesn't seem to be any such relationship as the distribution of values when overlaid by neighborhoods are almost identical to each other. As we expect, the higher rated neighborhoods have a lower mean PM2.5 concentration level, but this difference is less than 0.5 between A and D neighborhoods. Furthermore, since the differences between mean PM2.5 concentration values for each of the neighborhoods are so small, they would not be good predictors of racist neighborhood homeowner loan corporation grades.

Multiple hypothesis testing/decision making

Methods

The null hypotheses that are tested are listed as follows with each being referenced with its test number:

1. There is no difference in the *ozone concentration* levels between neighborhoods with an A grade and neighborhoods without an A grade.
2. There is no difference in the *ozone concentration* levels between neighborhoods with a B grade and neighborhoods without a B grade.
3. There is no difference in the *ozone concentration* levels between neighborhoods with a C grade and neighborhoods without a C grade.
4. There is no difference in the *ozone concentration* levels between neighborhoods with a D grade and neighborhoods without a D grade.
5. There is no difference in the *PM2.5 concentration* levels between neighborhoods with an A grade and neighborhoods without an A grade.
6. There is no difference in the *PM2.5 concentration* levels between neighborhoods with a B grade and neighborhoods without a B grade.
7. There is no difference in the *PM2.5 concentration* levels between neighborhoods with a C grade and neighborhoods without a C grade.
8. There is no difference in the *PM2.5 concentration* levels between neighborhoods with a D grade and neighborhoods without a D grade.

A multiple hypothesis test arises here because from a singular dataset, there are multiple questions. There were eight null hypotheses, one for each of the eight hypothesis tests that were conducted. It makes sense to test many hypotheses rather than just one because there are four different grades that are given to the various neighborhoods and two different metrics of evaluation, namely the ozone concentration level and the PM2.5 concentration level. Separately testing metrics by grade level makes for a clearer and more robust comparison of concentrations and reduces the probability of falsely rejecting tests.

For each hypothesis test, an A/B test is conducted by shuffling whether or not a certain neighborhood is in a particular grade category. Binary columns are added to the dataset to indicate a given neighborhood's grade. For example, if a given neighborhood is assigned the grade A, in the column 'A,' the value will be marked as True. In the columns 'B,' 'C,' and 'D,' the values will be marked as False. A/B tests are used here because for each of the hypotheses, two clear dataset groups formulate based on the categorical grade variable. Then, the relationship between these two groups is tested to see whether or not there is a statistically significant difference which would indicate that there is a difference in a certain evaluation metric between groups with a certain grade and groups without a certain grade.

We use Bonferroni correction and the Benjamini-Hochberg procedure to correct for the multiple hypothesis tests. Bonferroni correction controls the Family Wise Error Rate (FWER) which is the probability that any of the eight hypothesis tests conducted is a false positive. Benjamini-Hochberg controls the False Discovery Rate (FDR), which is the False Discovery Proportion (FDP) averaged over the randomness in the sequence of p-values from the hypothesis tests. The p-value that has been set for if we were merely conducting one hypothesis test is p = 0.05. This is the limit for the False Positive Rate (FPR) for the singular test. However, since multiple tests are being conducted, this increases the chance of seeing false positives; hence, the application of controlling measures.

Results

Individual Test Performance

1. Test 1, 2, 3, 7, 8
    a. Result
        i. The p-values for the respective tests of 1, 2, 3, 7, 8 are 1.0000, 0.9398, 0.1366, 0.9999, 0.9999
    b. Interpretation
        i. These p-values indicate that we **don't** reject the null hypothesis, meaning there is no difference in either the PM2.5 or ozone concentration level between A and non-A grade neighborhoods depending on the test. Hence, we can determine that there is not a statistically significant relationship based on the p-value threshold value of 0.05. If there is no difference, there is no correlation with redlining practices, as higher differences in concentration levels would serve as evidence for redlining. Furthermore, it is important to note that such a high p-value could be because of the data and our binary columns.
2. Test 4, 5, 6
    a. Result
        i. The p-value for the respective tests of 4, 5, 6 are 0.00, 0.00, and 0.00.

b. Interpretation

i. These p-values of 0.000 indicate that we **do** reject the null hypothesis, meaning there is no difference in either the PM2.5 or ozone concentration level between A and non-A grade neighborhoods depending on the test. Hence, we can determine that there is a statistically significant relationship based on the p-value threshold value of 0.05. Additionally, A grade neighborhoods are known to be the best rated, so any difference in concentration levels preserves the higher quality of the neighborhoods. We can determine that redlining practices are in effect and external factors are at play such as factories dumping chemical waste near there. However, additional analysis must be conducted to determine why the p-value is 0. At initial glance, our code is binary and the data does not have huge variations.

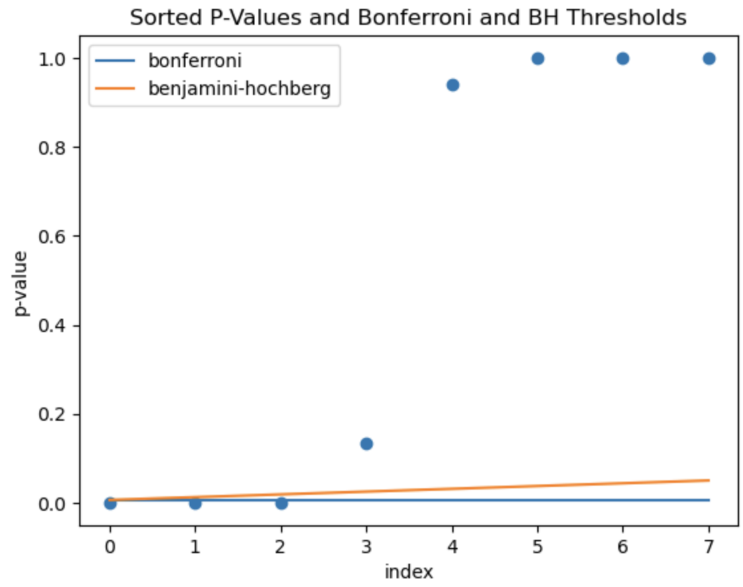


Figure 4: Graph showing p-values from eight hypothesis tests in order and the Bonferroni and Benjamini-Hochberg “new p-value” thresholds for each index

Bonferroni Correction

Since the original α = 0.05, for a singular test, for Bonferroni correction, $\frac{0.05}{8}$ = 0.00625, was used as the new threshold for all of the eight tests above. This resulted in rejecting hypotheses with p-values ≤ 0.00625. The hypotheses rejected were:

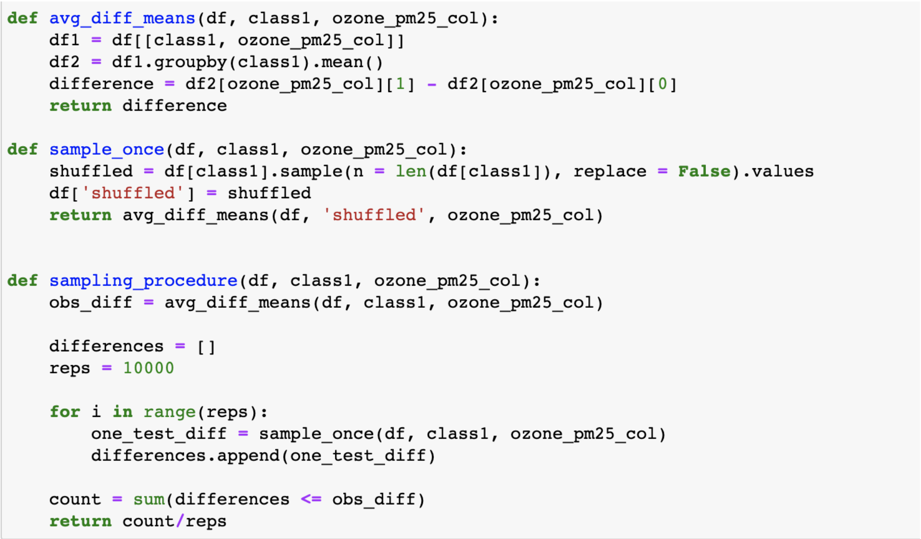
- Test 4 (hypothesis above)
- Test 5 (hypothesis above)
- Test 6 (hypothesis above)

Benjamini-Hochberg

To apply the Benjamini-Hochberg procedure, we first sorted all of the p-values above from the eight tests in ascending order to get the following list,

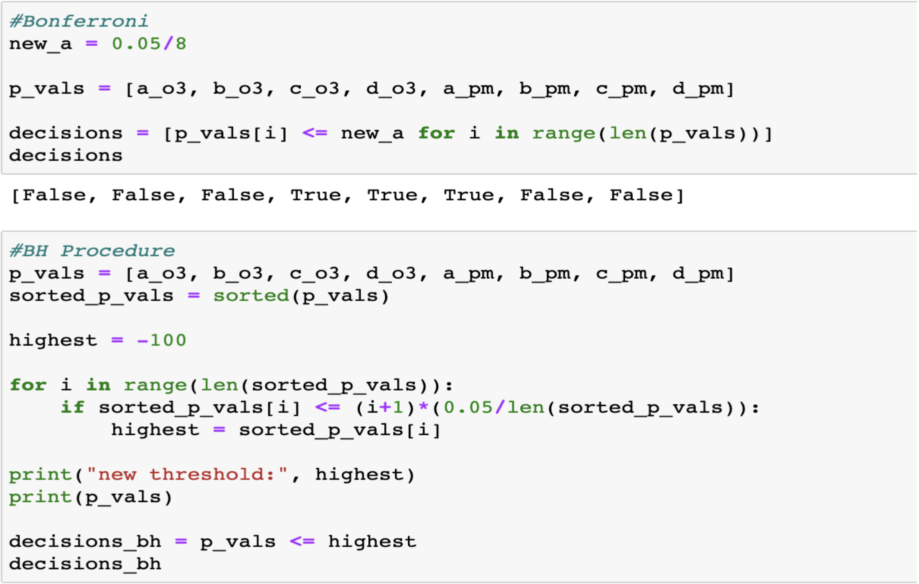
$[0.0000,\ 0.0000,\ 0.0000,\ 0.1288,\ 0.9427,\ 0.9999,\ 1.0000,\ 1.0000]$. The orange line in Figure 4 is drawn using the equation $y = i \cdot \frac{\alpha}{m}$, where i is the index of each value in the sorted p-values array starting with 1, α = 0.05, and $m = 8$ to indicate the number of tests conducted. After plotting this line, we found the largest p-value under this line to be 0. This then became the new p-value threshold with which to compare with the original eight p-values from the tests above. Using this new threshold, the hypotheses rejected were:

- Test 4 (hypothesis above)
- Test 5 (hypothesis above)
- Test 6 (hypothesis above)



```python
def avg_diff_means(df, class1, ozone_pm25_col):
    df1 = df[[class1, ozone_pm25_col]]
    df2 = df1.groupby(class1).mean()
    difference = df2[ozone_pm25_col][1] - df2[ozone_pm25_col][0]
    return difference

def sample_once(df, class1, ozone_pm25_col):
    shuffled = df[class1].sample(n = len(df[class1]), replace = False).values
    df['shuffled'] = shuffled
    return avg_diff_means(df, 'shuffled', ozone_pm25_col)


def sampling_procedure(df, class1, ozone_pm25_col):
    obs_diff = avg_diff_means(df, class1, ozone_pm25_col)

    differences = []
    reps = 10000

    for i in range(reps):
        one_test_diff = sample_once(df, class1, ozone_pm25_col)
        differences.append(one_test_diff)

    count = sum(differences <= obs_diff)
    return count/reps
```

Figure 5: Code applying techniques from class for methods for applying A/B test including finding the average difference in means, shuffling labels, bootstrap, and sampling repeatedly



```
#Bonferroni
new_a = 0.05/8

p_vals = [a_o3, b_o3, c_o3, d_o3, a_pm, b_pm, c_pm, d_pm]

decisions = [p_vals[i] <= new_a for i in range(len(p_vals))]
decisions
```

```
[False, False, False, True, True, True, False, False]
```

```
#BH Procedure
p_vals = [a_o3, b_o3, c_o3, d_o3, a_pm, b_pm, c_pm, d_pm]
sorted_p_vals = sorted(p_vals)

highest = -100

for i in range(len(sorted_p_vals)):
    if sorted_p_vals[i] <= (i+1)*(0.05/len(sorted_p_vals)):
        highest = sorted_p_vals[i]

print("new threshold:", highest)
print(p_vals)

decisions_bh = p_vals <= highest
decisions_bh
```

Figure 6: Code detailing how decisions were obtained after controlling for FWER and FDR using Bonferroni correction and the Benjamini-Hochberg procedure

Bonferroni correction controls for the Family Wise Error Rate (FWER). The Benjamini-Hochberg Procedure controls for the False Discovery Proportion (FDR). As seen above, the new threshold from Bonferroni correction reduced the p-value threshold to limit false positives. The Benjamini-Hochberg procedure also reduced the p-value threshold to limit the FDR.

Discussion

The statistically significant discoveries were in tests 4, 5, and 6. A description of these tests is listed above, and these discoveries mean that there is a difference in either the ozone or PM2.5 concentration levels between the specified neighborhood and the other neighborhoods.

From the individual tests, some key decisions that should be made are that racist redlining practices are not in effect if the null is not rejected. For instance, if there are no differences in concentration levels, policy groups and the government shouldn't have to investigate further if there are redlining practices in effect. In aggregate, we can see that the majority of tests don't seem to reject the null hypothesis, so there is less evidence to support a correlation between redlining and concentration levels.

Limitations in this analysis include that we solely looked at the main class that was assigned to each neighborhood. Each neighborhood was also assigned a secondary and tertiary class. However, these were not included in the hypothesis tests that were conducted. Relationships between the primary, secondary, and tertiary classes were also not included in the analysis.

P-hacking was avoided by using Bonferroni correction and the Benjamini-Hochberg procedure. These adjustments prevented potential incorrect discoveries that arise just by the nature of conducting multiple hypothesis tests. Moreover, we replicated the results by conducting the hypothesis tests several times first without setting a random seed.

If there was more data, additional tests that would have been conducted include correlation tests between more specific pollutant concentrations measured in the various neighborhoods. Correlation tests would have allowed us to conclude whether or not there are statistically significant correlations between certain pollutants and PM2.5 and ozone concentrations. This would strengthen our overall analysis.

Prediction with GLMs and Nonparametric methods

Methods

We use the average ozone and PM2.5 concentrations from 2011-2014 and their standard errors to see if redlining practices are biased. To do this we simplify the four letter grades into a single binary feature 'affluent' which is 1 if a neighborhood grade is A or B, and 0 if the grade is C or D. This allows the models to be less sensitive to changes in concentration levels. We also added linear combinations of ozone and PM2.5 concentrations as features. This helps our models as it creates more data to train on. A final new feature we created is 'high_pm' which is 1 if PM2.5 concentrations are above 12 and 0 otherwise, since healthy PM2.5 levels range between 0-12.

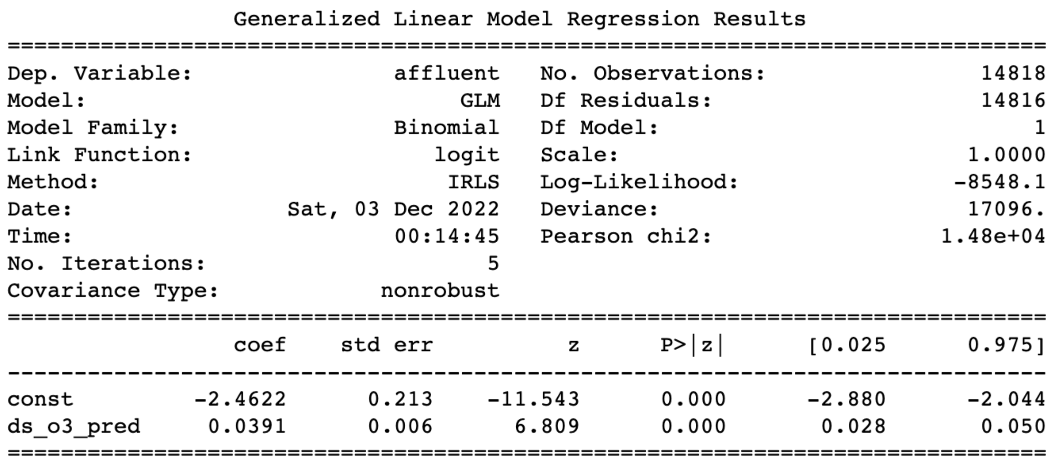
We began with a Frequentist GLM, which does not require assumptions regarding the prior distributions, and we used logistic regression since we aim to classify neighborhoods. Hence, the inverse link function is sigmoid, and the link function is logit with a Binomial likelihood, as an extension of the Bernoulli distribution. Our independent variables are ozone and PM2.5 and the dependent variable is the binary 'affluent.' We then implemented a Bayesian GLM using the same variables. We assumed that the affluence of a neighborhood, i.e. our θ, is random. We define the distribution of θ to be a Binomial distribution. For the initial Bayesian GLM, since we did not have extensive prior knowledge, we took a generalized approach and assumed the α and β parameter distributions to be normal.

The nonparametric methods used were a random forest and a deep neural network (scikit-learn's Multilayer Perceptron classifier) as both do not make prior assumptions about the data and perform well for a variety of features, numerical and binary. We will use the accuracy score on a held out test set of the data. For the Frequentist GLM, we will evaluate the model's performance based on the log likelihood, Pearson Chi-squared metric, and a comparison with a bootstrap inference of the data. For Bayesian GLM, we will look at the accuracy of the model along with coefficient and intercept traces.

Results

Frequentist Models:

Figure 7 shows the GLM model result for affluent predictions using ozone values, and Figure 10 shows the GLM model results for affluent predictions using PM2.5 values. For the frequentist models, to estimate uncertainty in the GLM predictions we used bootstrapping to randomize various intercept values for runs of the binomial model to approximate the sampling distribution as shown in Figure 8 and 11. We find that using the likelihood values, predicting neighborhood grades with PM2.5 is a slightly better fit than using the ozone because the log likelihood value is slightly larger. However, the values for both models are large, and don't provide much other useful information.



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:               affluent   No. Observations:                14818
Model:                            GLM   Df Residuals:                    14816
Model Family:                Binomial   Df Model:                            1
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -8548.1
Date:                Sat, 03 Dec 2022   Deviance:                       17096.
Time:                        00:14:45   Pearson chi2:                 1.48e+04
No. Iterations:                     5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -2.4622      0.213    -11.543      0.000      -2.880      -2.044
ds_o3_pred     0.0391      0.006      6.809      0.000       0.028       0.050
==============================================================================
```

Figure 7: Results from statsmodel GLM (logistic regression) for ozone predictions

Looking at Figure 7, we find that the ozone coefficient value is 0.0391. This means that holding everything else constant, every one unit increase in ozone concentration increases the log odds ratio of being in an affluent neighborhood by 0.0391. The affluent group has $e^{0.0391} \approx 1.04$ times the odds of the non-affluent group. This result is small, so there doesn't seem to be a drastic odd in ozone being a predictor of neighborhood affluence based on redline grade levels. Moreover, our confidence intervals are narrow, so the model is sure of this result.

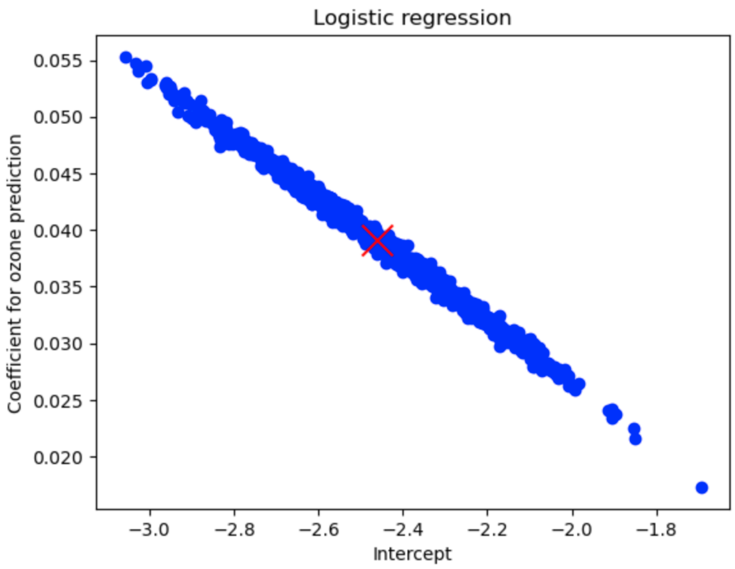
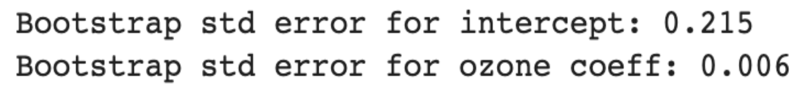


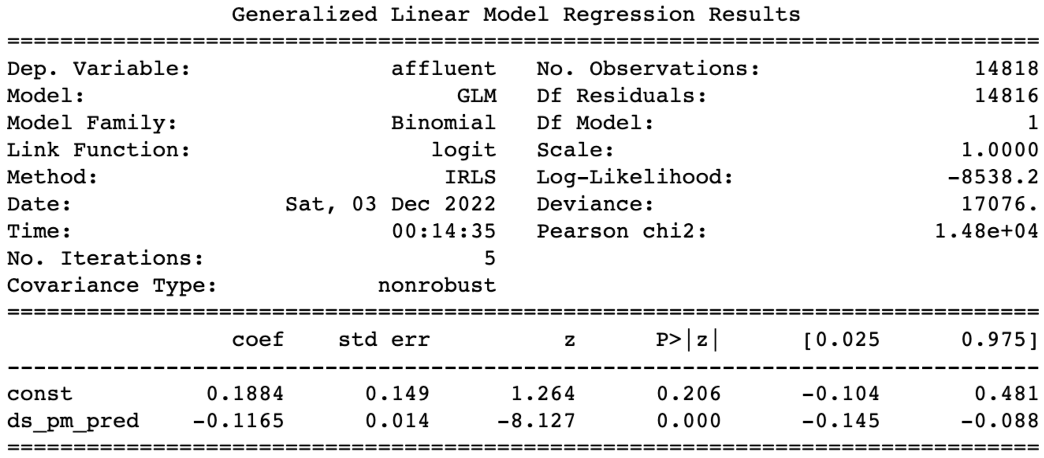
Figure 8: Bootstrap distribution for ozone predictions



```
Bootstrap std error for intercept: 0.215
Bootstrap std error for ozone coeff: 0.006
```

Figure 9 : Bootstrap error rates for ozone predictions

Figure 8 shows the bootstrap distribution. Looking at the x-axis (intercept) coefficients, the possible values are of a wider range, so there is more uncertainty present in the data. However, the standard errors (Figure 9) are actually the same as that of the GLM model, so it is uncertain but not a bad predictor. The bootstrap error is larger for the intercept, so it seems that our model is fairly good, but there is some small uncertainty.



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:               affluent   No. Observations:                14818
Model:                            GLM   Df Residuals:                    14816
Model Family:                Binomial   Df Model:                            1
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -8538.2
Date:                Sat, 03 Dec 2022   Deviance:                       17076.
Time:                        00:14:35   Pearson chi2:                 1.48e+04
No. Iterations:                     5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.1884      0.149      1.264      0.206      -0.104       0.481
ds_pm_pred    -0.1165      0.014     -8.127      0.000      -0.145      -0.088
==============================================================================
```

Figure 10: Results from statsmodel GLM (logistic regression) for PM2.5 predictions

Looking at Figure 10, we find that the PM2.5 coefficient value is -0.1165. This means that holding everything else constant, every one unit increase in PM2.5 concentration decreases the log odds ratio of being in an affluent neighborhood by -0.1165. This means that PM2.5 concentrations are associated with a odds ratio of $1 - e^{-0.1165} \approx 0.11$ reduction in the probability of being in an affluent neighborhood. This is a somewhat significant change and demonstrates that PM2.5 might be a better predictor of affluence in neighborhoods compared to ozone. The confidence intervals are quite narrow, so the model is confident in its predictions.

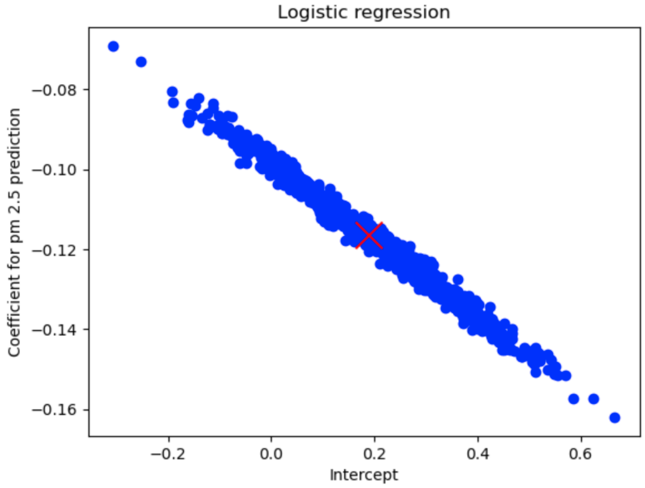
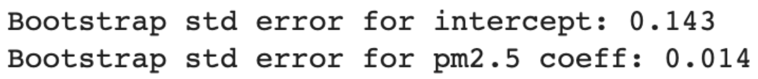


Figure 11: Bootstrap predictions for PM2.5 predictions



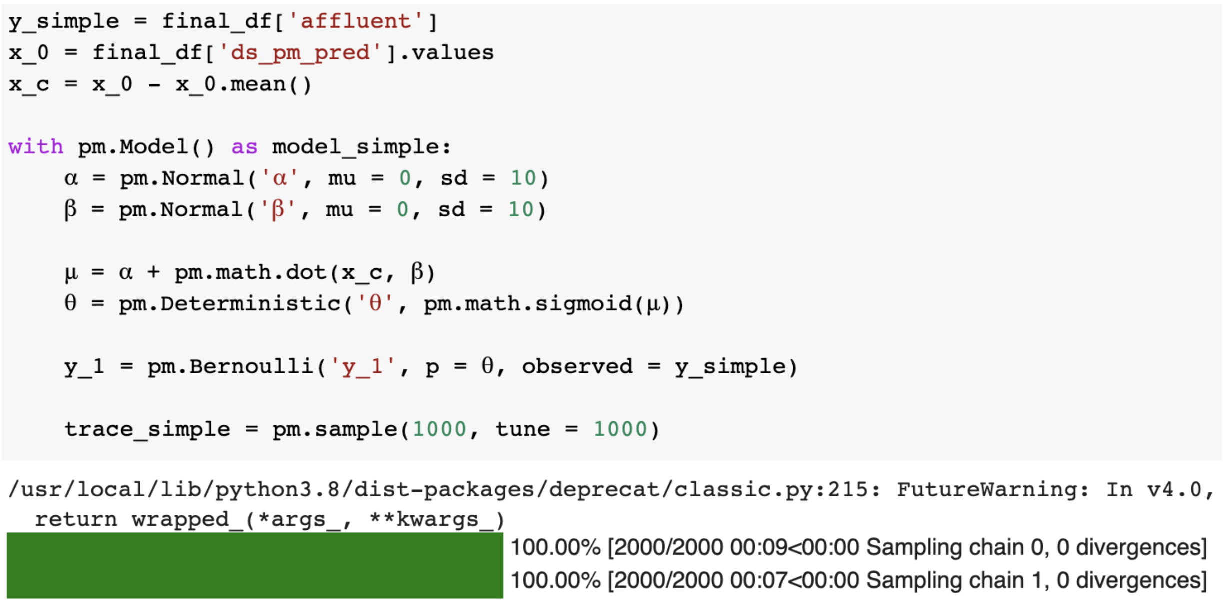
```
Bootstrap std error for intercept: 0.143
Bootstrap std error for pm2.5 coeff: 0.014
```

Figure 12: Bootstrap error rates for PM2.5 predictions

Figure 11 shows us the bootstrap distribution. Looking at the x axis (intercept) coefficients, the possible values are of a wider range, and there is more uncertainty in the data. Hence, the GLM was slightly overfitting and overconfident with its predictions. However, the range is not too different from the confidence interval of the model, and performs better than the ozone predictor. It seems that PM2.5 performs better at predicting neighborhood grades. The standard errors (Figure 12) are also smaller, which indicates confidence and decreased uncertainty.The bootstrap distribution also appears to predict less uncertainty than the GLM.

Bayesian Models:

We began with a relatively simple posterior predictive check on the data. First, we centered the PM2.5 data to help with the sampling. We used one of the deterministic variables θ as the output of the logistic function applied to the μ variable. The sigmoid function can be interpreted as whether or not the neighborhood is considered an affluent neighborhood, given that we know the PM2.5 concentrations. The code below shows this process:

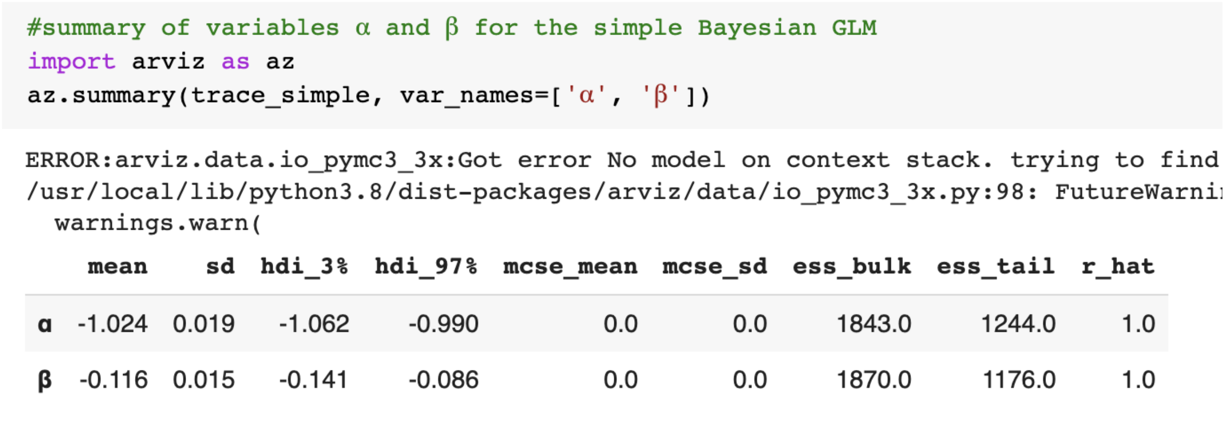


```
y_simple = final_df['affluent']
x_0 = final_df['ds_pm_pred'].values
x_c = x_0 - x_0.mean()

with pm.Model() as model_simple:
    α = pm.Normal('α', mu = 0, sd = 10)
    β = pm.Normal('β', mu = 0, sd = 10)

    μ = α + pm.math.dot(x_c, β)
    θ = pm.Deterministic('θ', pm.math.sigmoid(μ))

    y_1 = pm.Bernoulli('y_1', p = θ, observed = y_simple)

    trace_simple = pm.sample(1000, tune = 1000)
```

```
/usr/local/lib/python3.8/dist-packages/deprecat/classic.py:215: FutureWarning: In v4.0,
  return wrapped_(*args_, **kwargs_)
100.00% [2000/2000 00:09<00:00 Sampling chain 0, 0 divergences]
100.00% [2000/2000 00:07<00:00 Sampling chain 1, 0 divergences]
```

Figure 13: Simple PM2.5 predictive model using sigmoid function

We then summarized the inferred parameter values for an easier analysis of the results and to further understand how well the model did. The overall distribution of α and β parameters is normal and the distribution of y is Bernoulli. The α and β parameter values as well as their uncertainties, encoded with standard deviations, are in Figure 14. With this information, we ran a posterior predictive check on the simple GLM to explore how well our model captured the PM2.5 data. Results are shown in Figure 15, with an accuracy of 73%.

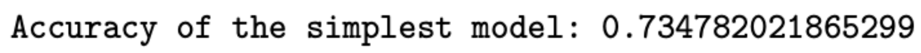


```
#summary of variables α and β for the simple Bayesian GLM
import arviz as az
az.summary(trace_simple, var_names=['α', 'β'])
```

```
ERROR:arviz.data.io_pymc3_3x:Got error No model on context stack. trying to find
/usr/local/lib/python3.8/dist-packages/arviz/data/io_pymc3_3x.py:98: FutureWarni
  warnings.warn(
```

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
α	-1.024	0.019	-1.062	-0.990	0.0	0.0	1843.0	1244.0	1.0
β	-0.116	0.015	-0.141	-0.086	0.0	0.0	1870.0	1176.0	1.0

Figure 14: Results for parameter values from simple logistic model



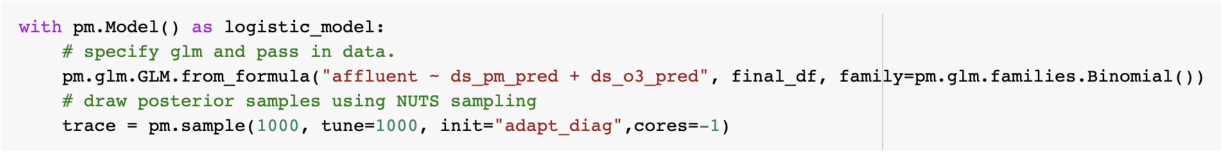
```
Accuracy of the simplest model: 0.734782021865299
```

Figure 15: Accuracy after running the simple logistic model

From here, we coded a slightly more advanced GLM that assumed that the affluence of a neighborhood is a function of PM2.5 and Ozone concentrations. In other words, the formula would be:

$$logit = \beta_0 + \beta_1(PM2.5) + \beta_2(Ozone)$$

Our goal was to determine whether a neighborhood is considered affluent or not based on the PM2.5 and ozone concentrations, so we used PyMC3 to draw samples from the posterior. We added the aforementioned variables to the model and generated a trace plot visible in Figure 17. The trace plot displays all of the samples drawn for all of the variables. The left shows the final approximate posterior distribution for the model's parameters, and the right shows the individual sampled values at each step during the sampling process. This GLM model behaves similarly to the simple model we created earlier. The model renders an accuracy of about 73% as shown in Figure 18, which again is quite similar to our previous model.



```
with pm.Model() as logistic_model:
    # specify glm and pass in data.
    pm.glm.GLM.from_formula("affluent ~ ds_pm_pred + ds_o3_pred", final_df, family=pm.glm.families.Binomial())
    # draw posterior samples using NUTS sampling
    trace = pm.sample(1000, tune=1000, init="adapt_diag",cores=-1)
```

Figure 16: Advanced Logistic model

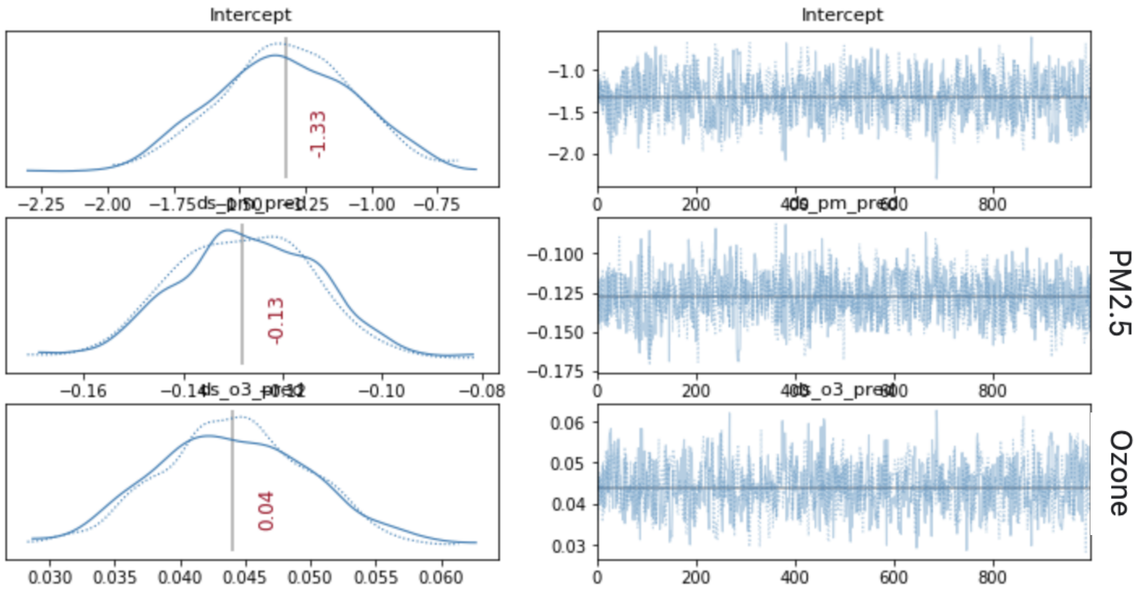
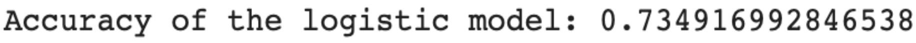


Figure 17: Trace plot to determine intercepts and coefficients, i.e. *β0, β1(PM2.5), β2(Ozone)*



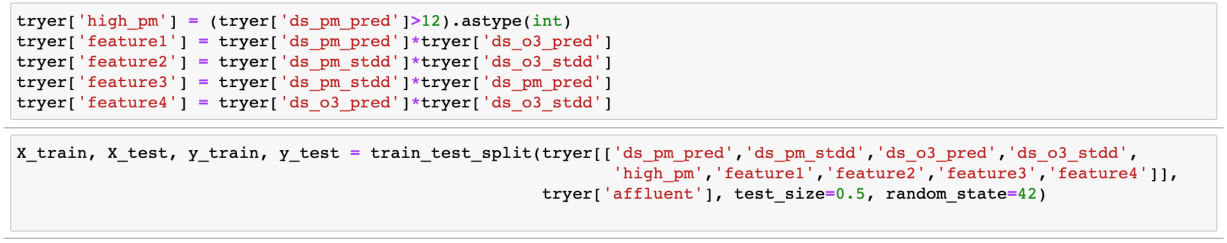
```
Accuracy of the logistic model: 0.734916992846538
```

Figure 18: Accuracy after running the advanced logistic model

From the graphs in Figure 17, we are able to see what the GLM outputs as the coefficients for the two variables and the value of the intercept. The uncertainty of each of these values is encoded within the distributions of the potential values that each can take on along with their respective probabilities. The intercept of -1.33 (β_0) tells us the estimated value of the response variable (affluence) when the continuous explanatory variables (PM2.5 and Ozone) have a value of 0. Furthermore, there are the coefficients of -0.13 (β_1) and 0.04 β_2) describing the slope of the model's predicted relationship between the PM2.5/Ozone concentrations and neighborhood affluence. A slight increase in the ozone concentration would contribute to an increase in the chance of being affluent, and a slight decrease in the PM2.5 concentration would contribute to an increase in the chance of being affluent. Of course, these must also vary simultaneously to create the seen effect.

Random Forest Model:

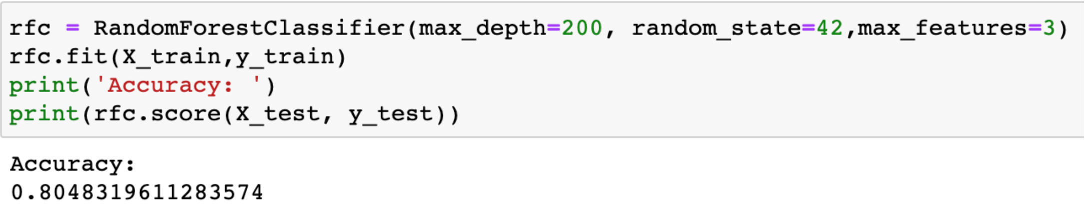
In order to maximize the random forest model, we first split up the data into a train and test set as shown in Figure 19. The obtained accuracy is approximately 0.80. This means that for each of the ozone and PM2.5 concentration level values, the model can accurately predict whether or not this concentration is from an affluent neighborhood or not 80% of the time. This is quite a high accuracy, and definitely better than random. This leads us to believe that ozone and PM2.5 concentration levels are predictive of neighborhood redlining grade levels.



```
tryer['high_pm'] = (tryer['ds_pm_pred']>12).astype(int)
tryer['feature1'] = tryer['ds_pm_pred']*tryer['ds_o3_pred']
tryer['feature2'] = tryer['ds_pm_stdd']*tryer['ds_o3_stdd']
tryer['feature3'] = tryer['ds_pm_stdd']*tryer['ds_pm_pred']
tryer['feature4'] = tryer['ds_o3_pred']*tryer['ds_o3_stdd']
```

```
X_train, X_test, y_train, y_test = train_test_split(tryer[['ds_pm_pred','ds_pm_stdd','ds_o3_pred','ds_o3_stdd',
                                                           'high_pm','feature1','feature2','feature3','feature4']],
                                                    tryer['affluent'], test_size=0.5, random_state=42)
```

Figure 19: Feature engineering and train/test split of data



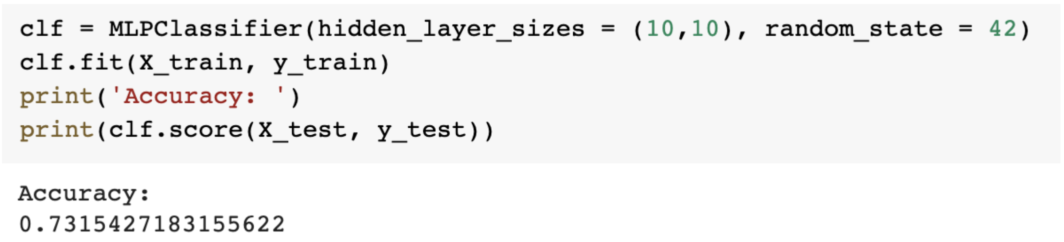
```
rfc = RandomForestClassifier(max_depth=200, random_state=42,max_features=3)
rfc.fit(X_train,y_train)
print('Accuracy: ')
print(rfc.score(X_test, y_test))
```

```
Accuracy:
0.8048319611283574
```

Figure 20: Results from Random Forest model

Neural Network Model:

Figure 21 below further shows the neural network code as well as the model's accuracy on the test data. We used the same train/test split as above (Figure 19) and found the accuracy to be approximately 0.73. Although not as high as the random forest, an accuracy of 73% is still very telling of the general trends we have seen between PM2.5/ozone concentrations and redlining grade levels. Hence, this further affirms that ozone and PM2.5 concentrations are predictive of affluent neighborhoods.



```
clf = MLPClassifier(hidden_layer_sizes = (10,10), random_state = 42)
clf.fit(X_train, y_train)
print('Accuracy: ')
print(clf.score(X_test, y_test))
```

```
Accuracy:
0.7315427183155622
```

Figure 21: Results from Neural Network model

Discussion

The nonparametric random forest model seemed to perform the best, with an advanced Bayesian logistic model coming in with second highest accuracy. This makes sense given that the random forest model has more features available to it than the GLMs, and also is able to utilize cross validation along with random feature selection for optimal results. It also makes sense that the Bayesian model performs better since we prespecify what the prior relationship is. There is a possibility that the random forest model may be overfitting to current data, and may perform worse for future data. The Bayesian model would be a better predictor given the relationship takes into account the prior distribution as a whole rather than specific data. The main differences observed between the Bayesian and Frequentist models are difficult to interpret given the Bayesian model combines PM2.5 and ozone and the Frequentist model separates them.

A limitation to the Frequentist GLM is similar to that of the random forest, since no assumptions are made, the GLM may inaccurately assume certain distributions, or overfit the data. A limitation to the Bayesian GLM is that the prior could be inaccurate to some degree, given that we do not have much field experience in this domain. Furthermore, if the posterior distributions are heavily affected by the prior, this can further prove to be a limitation in the model accuracy. Lastly, for the neural network, a limitation is the model explainability; neural networks are hard to explain. Additional data that could increase the accuracy of our models and help us understand if there is a correlation between neighborhood grades and pollution includes more data on the race demographics for each neighborhood grade and more greenhouse concentrations so we can measure pollution in addition to PM2.5 and ozone, as well as the incorporation of data on disease rates per region to further investigate the impact of pollution in each neighborhood grade and how that might be racially driven.

Conclusion

Through our analysis, we discovered that ozone and pollutant concentrations are slightly correlated with redlined grades, but fairly insignificant. Upon further analysis, there does not seem to be a strong correlation present in each grade level. However, our algorithm produced predictions of neighborhood grades better than random from ozone and PM2.5 concentration levels. Our findings are generalizable and focus on the entirety of the United States. We also find that even with the slight correlations, pollutant concentrations still predict neighborhood grades with better than random accuracy. Although our research hasn’t strongly proven that redlining practices are racist and that there must be action taken to address this, our research can be used as a guiding factor to further research and investigate the slight correlation we found between pollution and redline grades.

For further research, it would be interesting to study specific states in more detail with up to date information, or possibly places outside of America. Since we merged three different datasets (ozone concentrations, PM2.5 concentrations, and redlined grades), we lost specific temporal data because we grouped each dataset by city codes. However, by combining the datasets, we were able to extract meaningful insights when using the concentration levels of PM2.5 and ozone for various predictions. Some limitations to our research are that neighborhood redlining grades were produced decades ago, so neighborhood demographics, etc may have significantly changed. Another research question to follow up on is to study the effect of greenhouse gas emissions in relation to neighborhood grades and use more recent neighborhood sentiment. Environmental agencies can use these results to take legal action/suggest financial penalties against companies polluting these areas.

References

Home Owner Loan Corporation (HOLC) neighborhood grades for US Census tracts - home owner Loan Corporation (HOLC) neighborhood ratings for 2010 census tracts - CKAN. Home Owner Loan Corporation (HOLC) neighborhood grades for US census tracts - Home Owner Loan Corporation (HOLC) neighborhood ratings for 2010 census tracts - CKAN. (n.d.). Retrieved December 9, 2022, from https://data.diversitydatakids.org/dataset/holc_census_tracts-home-owner-loan-corporation--holc--neighborhood-grades-for-us-census-tracts/resource/848dbd07-6913-4757-85bc-fc0c37f5c6fb