HON 4355

FINAL REPORT

ANAND SABNIS

A diagram of a network

Description automatically generated2055017

APPENDIX

1. Motivation
2. Data and Metadata
3. Analysis
4. Results and Interpretation

**Motivation**

The discussion of your motivation should be as detailed as feasible, based on considerations we

discussed in class.

At a minimum, describe your chosen question and your reasoning for:

- How answering your question contributes towards solving a problem (e.g., it shows how

widespread the problem is, or how persistent the problem is, or how effective a possible

solution to the problem is, or how ineffective a current mitigation effort of the problem

is, or how closely related to/dependent on another issue the problem is), and

- Why this problem matters for society (e.g., what is the cost to society of leaving the

problem unsolved, what are the possible gains if the problem was solved).

Disciplining in schools has been something that I’ve researched in the past. When I first started research, disciplining seemed very one-dimensional, with a very simple concept- “misbehave and be disciplined”. However, after learning the importance of causal analysis, I’ve realized that this is a special case. Throughout the semester, I’ve worked towards understanding the ‘why’ with this research. Like any other Data and Society problem, this one is multifaceted with no simple fix. These problems affect us at large and have multiple causes and effects. This is why bringing about change is tough.

I’ve looked at disciplining in Houston in particular. Last semester, I was casually browsing all the ISDs when I decided to see the stats for the lowest-ranked ISD. It was Hempstead ISD. This ISD consisted of 4 elementary schools, 3 middle schools, and 1 high school. I investigated Hempstead town and came across a real estate website called ‘area vibes’. It showed me that the unemployment rate and poverty levels were way above the state average. At the same time, the real estate prices, along with median income were below average. Further research into the history of the town showed that:

Most of the people in the town are self-employed, i.e. own small retail businesses. They are followed by people who work in manufacturing- a local automobile plant employed a couple of hundred workers. Unfortunately, the plant closed a few years ago, and the workers have had to migrate to other jobs. Due to being unskilled, they are having a hard time finding a stable job. The neighborhood has high crime, and this in addition to a scarcity of jobs has driven real-estate prices down the drain.

The effects of these adverse conditions were felt inside schools too. A huge majority of Hempstead High School students were at-risk of dropping out, were poor, and had limited English proficiency.

At the time, I was focused on understanding the effects of neighborhood factors on in-school statistics. However, I couldn’t find a good way to expand on this and get actionable data for multiple schools. The highly qualitative data stopped me from making a quantitative analysis.

This is when I came across a research paper that was part of the RICE HERC. The article talked about how prior out-of-school suspensions had a strong linear relationship with future juvenile justice contact, and once these students come back to school, the chances of future contact are higher. (Duffy et al. 2021) Contrary to common beliefs, students do not have contact during the suspension period but over time. On average, in HISD, Black, Latinx and economically disadvantaged students have higher odds of receiving an OSS.(Duffy et al. 2021)

In short, getting even a single OSS can lead to having 6 times more susceptibility of facing future juvenile justice contact. (Duffy et al. 2021) The work that Dr. Duffy did mainly pertained to chances of getting OSS and then the chances of receiving future OSS.

I was interested in doing similar research that would allow me to understand what factors affect the number of OSS the most.

My research aims to address the following questions:

1. Does the number of OSS vary by student characteristics?

2. What explains the relationship between suspensions and juvenile justice contact?

**DATA AND METADATA**

The discussion of your data and metadata should be as detailed as feasible, based on

considerations we discussed in class (e.g., datasheets for datasets). Use enough detail so that

someone else could replicate your work and reproduce the dataset you are working with (this is

meant to be in principle – I will not actually try to replicate your work).

At a minimum, describe:

- The source(s) you used to compile your dataset (e.g., actual URL, file name),

- The variables you decided to include in your dataset (e.g., name, meaning), and

- Any filtering/selection processes that resulted in modifying the above by

excluding/dumping/leaving out data.

Also provide your reasoning that led to your specific methodological choices (i.e., the decisions

you made for where and how to collect and compile your dataset). You could consider providing

a link to a shared repository where your actual dataset is hosted, or even include it as an appendix

itself. Also reference the previous work that helped you formulate your reasoning (by citing references

where appropriate). Remember, references are always the first appendix in a research report.

**The datasheet for both the datasets (with all variables) can be found here:**

**https://drive.google.com/drive/folders/1VbVfqWAGbdwl28HUV9V7DSmqfelB0ekj?usp=sharing**

The filtering process for this dataset was complex and time-consuming.

This is step-by-step the filtering process that I adopted:

1. Remove all school districts that are not HISD: Since the campus summary covers all schools in Texas, I had to remove all schools that were not considered part of HISD. This involved selecting ‘Houston ISD’ from the CAMPUS NAME and COLUMN.
2. Next came picking High Schools specifically from HISD. For this I had to take 2 steps:
   1. Filter out elementary and middle schools using the keywords: ’Primary’ and ‘EL’.
   2. Go through the list manually and filter out the schools that were not high schools (this process was especially time consuming).

The schools that were removed using this process were:

BAYLOR COLLEGE OF MEDICINE BIOTECH 101912234

BAYLOR COLLEGE OF MEDICINE ACADEMY 101912467

ARABIC IMMERSION MAGNET SCHOOL 101912478

FARIAS EARLY CHILDHOOD CENTER 101912352

FONWOOD EARLY CHILDHOOD CTR 101912470

GARDEN OAKS MONTESSORI 101912157

GREGORY-LINCOLN ED CTR 101912058

H S AHEAD ACADEMY 101912456

HALPIN EARLY CHILDHOOD CTR 101912131

KING EARLY CHILDHOOD CTR 101912355

LAS AMERICAS 101912340

LAURENZO EARLY CHILDHOOD CTR 101912357

MANDARIN IMMERSION MAGNET SCHOOL 101912460

MISTRAL CENTER FOR EARLY CHILDHOOD 101912354

NEFF ECC 101912209

PILGRIM ACADEMY 101912218

REAGAN K-8 EDUCATIONAL CTR 101912382

SCHOOL AT ST GEORGE PLACE 101912353

SUGAR GROVE ACADEMY 101912163

TSU CHARTER LAB SCH 101912328

WHARTON K-8 DUAL LANGUAGE ACADEMY 101912256

WILSON MONTESSORI 101912259

WOODSON SCHOOL 101912127

1. Then I picked H-05-OUT-OF-SCHOOL SUSPENSIONS from the HEADING NAME column and extracted the counts for each high school in HISD.
2. I also extracted the total discipline counts and total enrollment counts for each high school and added them as a column.
3. I also extracted total state discipline counts for the years 2017-19, to help me build an argument.

My second dataset was custom and manually generated by me due to there being a lack of a ‘dataset’ that gave me the variables I wanted.

I used the USNews website to get counts of:

1)Racial breakdown(percentHispanic percentBlack percentWhite percentTwo or More Races percentAmerican Indian/Alaska Native percentAsian percentNative Hawaiian/Pacific Islander)

2) Gender(percentFemale, percentMale)

3) Econ. Disadv.(percentEconDisadv)

4) Student:Teacher Ratio

For the years 2017-19.

All of these became columns in my new custom dataset.

My dataset now had 12 columns and 46 rows.

I used Excel to compile, edit and clean all my datasets.

I used R to execute and make graphs for descriptive statistics and modeling

**Analysis**

Provide summary statistics for your data; they should be as detailed as feasible, based on

considerations we discussed in class. Use enough detail so that someone else could verify that

they are working with the same dataset as you (this is meant to be in principle – I will not actually

try to replicate your work).

At a minimum, compute and include descriptive statistics for single variables or pairs of variables

in the form of

- tables (e.g., medians, means, IQRs, standard deviations) and/or

- graphs (e.g., scatterplots).

Also provide your reasoning that led to your specific methodological choices (i.e., how you

decided which tables and/or graphs were important enough to include here).

Since my final dataset is custom made, this is what the first 20 rows of my dataset look like:

(Link to dataset: https://docs.google.com/spreadsheets/d/1f-MeH0yuYUS1LWhhnbj0uoLnonfM0zcE/edit?usp=sharing&ouid=101720594549894684356&rtpof=true&sd=true)

Again, the predictor variable here is percentOSS.

A white background with many small squares

Description automatically generated with medium confidence

A screen shot of a graph

Description automatically generated

This correlation matrix highlights the relationships between the response variable: percentOSS  
and predictor variables: percentHispanic, percentWhite, percentTwo or More Races, percentAmerican Indian/Alaska Native, percentAsian, percentNative Hawaiian/Pacific Islander, percentMale, percentEconDisadv., Student:Teacher Ratio.

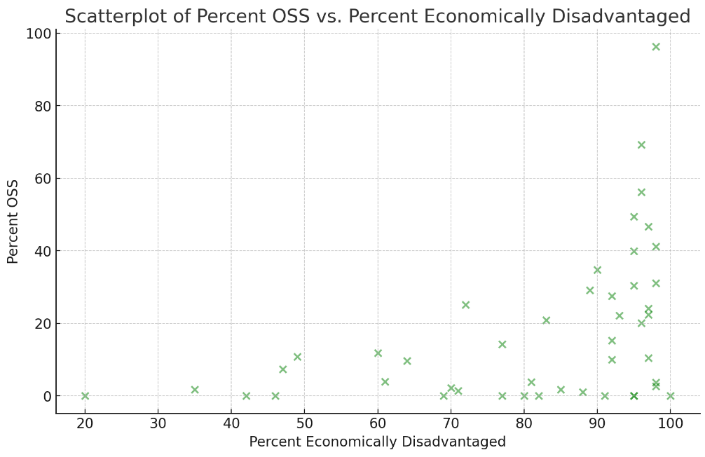
From the graph, the warmer colors indicate a positive relationship between 2 variables. For example, percentHispanic and percentWhite are moderately negatively correlated at -0.51. This means that they have an inverse relationship.

I added this plot first because it is a combination of multiple comparison graphs.

This plot shows variation in student populations across all schools. This graph will later be useful in understanding variations in data. For example, percentHispanic stayed consistent at >40% for most schools.

The percentBlack variable is more varied then percentHispanic. These 2 are the largest student bodies on average for all 46 schools in the study.

From the graph, they have a negative linear relationship.



The 2 scatterplots on the left highlight the relationship between percentOSS and percentEconDisadv. & percentOSS and student:teacher ratio(log of this function was used in final analysis).

1. The 2 variables OSS and econdisadv are linearly positively related.
2. The 2 variables OSS and Student:Teacher Ratio are strongly positively related.

A graph with blue dots

Description automatically generated

A screenshot of a graph

Description automatically generated

The pair plot on the left shows a mix of scatter plots and histograms.

Below the diagonal, there are scatter plots that show the relationship between two variables. Each point on a scatter plot represents an observation in the dataset with its respective values for the two variables being compared. For example, there's a scatter plot comparing "% Hispanic" on the x-axis and "OSS Avg." on the y-axis.

Above the diagonal, we see numerical values that represent the correlation coefficient between the variables of the corresponding row and column. The correlation coefficient is a statistical measure that describes the strength and direction of a relationship between two continuous variables. For eg., the correlation coefficient between "% Hispanic" and "OSS Avg." is -0.286, indicating a weak negative correlation.

All of these plots were picked because they talk about relationships that my model doesn’t.

In order to make my model fit and have a good AIC, I had to deal with overfitting and colliearity issues.

This led me to remove a lot of these variables from the final prediction model.

**Results and Interpretation**

Provide a detailed discussion of your process of analyzing the data and the conclusions this

analysis leads to. Use enough detail so that someone else could critically engage with your

argument (i.e., follow along, understand, and be able to offer relevant and useful suggestions for

your interpretation).

At a minimum, document and discuss:

- how you interpret your results and how does this interpretation lead you to answer your

question,

- what is the evidence to support your interpretation, based on your process for analyzing

your dataset (i.e., why should a neutral reader believe you), and

- how strong you feel this evidence is, based on your understanding of the limitations about

your process, or dataset, or other factors (i.e., is there room to strengthen your

argument).

Also provide your reasoning that led to your specific methodological choices (i.e., how you

decided which process(es) were appropriate to use for analyzing the data). When applicable, you

could consider providing a link to a shared repository where your actual code is hosted, or even

include it as an appendix itself.

My model is a prediction model that prediction the %OSS count per school in HISD.

I tried the following models:

1)Linear Regression

2)Poisson Distribution

3)Negative Binomal

But all of them gave me high error rates, AICs or were just a bad overall fit.

One of the main reasons why the data was ‘causing trouble’ was due to a large number of zeroes in the data.(I’ve talked about this before, the FERPA guidelines force the TEA to insert placeholder values when the counts are low).

I finally decided on using a Poisson(Zero-inflated) Count Model.

The excess zeros are generated by a separate process from the count values and that the excess zeros can be modeled independently. Thus, the zip model has two parts, a Poisson count model and the logit model for predicting excess zeros.

Though we can run a Poisson regression in R using the glm function in one of the core packages, we need another package to run the zero-inflated Poisson model. We use the pscl package.

Since zip has both a count model and a logit model, each of the two models should have good predictors. The two models do not necessarily need to use the same predictors.

Count data often use exposure variables to indicate the number of times the event could have happened. I’ve included a log modified version of the exposure variable into my model by using the offset() option.

Instructions to run the code are in this .readme file:

Based on the summary output for the zero-inflated Poisson model you have provided, the equations for the count model and the zero-inflation model can be written as follows:

Poisson Count Model Equation (for the expected count):

*Log(lambda)= -21.972884 + 0.156035(percentHispanic) + 0.174624(percentBlack) + 0.187193(percentWhite) + 0.185162(percentAsian) + 0.022302(percentMale) + 0.035775(percentEconDisadv) + log(Total Enrollment)*

Zero-Inflation Model Equation (for the probability of excess zeros):

*logit(p) = -16.25024 + 0.185271(percentHispanic) - 0.02114(percentBlack) + 0.31236(percentWhite) + 0.18065(percentAsian) - 0.20590(percentMale) + 0.02515(percentEconDisadv)*

In these equations:

- ( lambda) is the expected count of the response variable `OSS Avg.` from the Poisson component of the model.

- (p) is the probability of an observation having an excess zero count, predicted by the zero-inflation component of the model.

- The log link function in the Poisson count model takes the natural logarithm of the expected count.

- The logit link function in the zero-inflation model takes the natural logarithm of the odds of an excess zero.

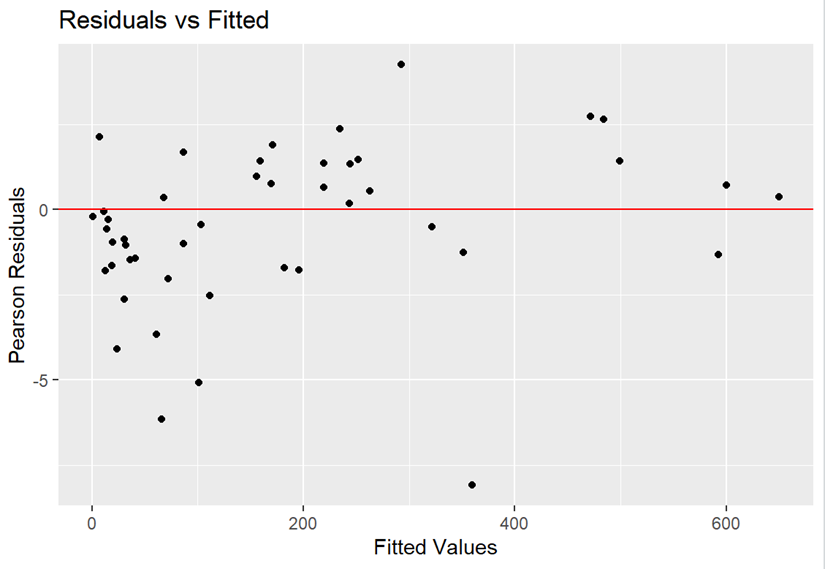
These equations are generated by exponentiating both sides of the equation for (lambda and ( p ) to get the actual expected count and probability. In the case of (p), you would use the logistic function to transform from log-odds to a probability.

Other variables were removed due to collinearity and overfitting issues.

There are 7 total variables used in the final model.

The results are interpreted(and shown in the section above) by looking at the estimate column and multiplying it by each predictor variable. Each estimate ranges from to and they acts as the weights that when multiplied by the counts, form the predicted variable.

The Resid. V/S Fitted Graph and the QQ Residuals Plot is as follows:

The line of best fit accommodates a moderate amount of values. However, this model is decent at best, as a significant portion of values do not fall on the model line.

A graph with a line

Description automatically generated

The QQ plot of residuals shows a good fit for most models. However, there is a significant dropoff on the left. This indicates that the model fails when the OSS count is significantly low.

The model summary is as follows:

A screenshot of a computer

Description automatically generated

**Conclusion**

Provide a reflection on your overall project. What are the key take away message(s)?

This does not need to be limited to your results and interpretation - it could be about any of the

work you did; what did you find out that you think would most benefit a reader to know?

This research gave me detailed insight into different factors that affect out-of-school suspensions in schools.

I was able to predict, with moderate certainty, what the OSS count for all schools in HISD will be when we know different factors.

The key takeaway messages are as follows:

1. Does the number of OSS vary by student characteristics?

Yes, it does. Infact, all student characteristics are significant in predicting the OSS counts.

1. What explains the relationship between suspensions and juvenile justice contact?

The presence of high OSS in schools points to a chance of the same students being disciplined more than once. One OSS incident increases chances of this contact exponentially.

Future Scope:

Researchers should try and look at qualitative data on schools, for example: Neighborhoods, Parental Involvement and Teacher Experience. This data won’t be actionable at higher levels until we do the grassroots level work first. To look at these variables, we need to start small-from a single school, and understand the problem.

This was my take on understanding disciplining.

I would like to propose some solutions to this problem, something that I’ve learned over the course of my time in Data and Society and my time spent learning about discipline:

These were a combination of my ideas, along with some pulled from the HISD code of conduct, as well as Rice University studies:

1. Deciding Priorities: Administrators should carefully review resource inequities when making budget decisions regarding allocations intended to improve school climate and security.
2. Restorative Justice: Discipline should be geared towards inclusivity and rehabilitating students back into the school system. Above all, students need to be stopped from dropping out.
3. Reducing Bias: Reducing bias will serve a dual purpose: Allow discriminated groups to have better futures after school, which will in turn boost school rankings, and eventually, help entire neighborhoods flourish.
4. Reducing Chances of future OSS: Implementing restorative justice will help us prevent students from falling in the cycle of disciplinary actions.

Breaking the school-to-prison pipeline needs to be our priority.

To conclude, making schools better serve as an ancestor of outcome. If we ensure our schools are geared towards keeping students in school, rather than forcing them to drop out, students will have better futures. The prison pipeline starts when school ends.

The more time we invest in our students, the less money will be spent keeping them in prison.

BIBLIOGRAPHY

1. Texas Education Agency (TEA). (n.d.). Discipline reports. Retrieved from <https://tea.texas.gov/reports-and-data/student-data/discipline-data-products/discipline-reports>
2. Rice University. (2020). School Suspensions Associated with Future Juvenile Justice Contact. Retrieved from <https://rice.app.box.com/s/cbhh5tkqud14qaujk4z1eq4jczx85dr7>
3. Rice University. (n.d.). The Consequences of Exclusionary School Discipline in Texas: The Impact of Suspension and Expulsion on Student Outcomes. Retrieved from <https://rice.app.box.com/s/cbhh5tkqud14qaujk4z1eq4jczx85dr7>
4. Rice University. (n.d.). Reversing the Pipeline: The Case for Culturally Responsive and Restorative Practices in Schools. Retrieved from <https://rice.app.box.com/s/xzuu1lcau9g62ik38ocht2zmi112dtjf>
5. Texas Criminal Justice Coalition. (2020). Reversing the Pipeline: A Comprehensive Approach to School Discipline Reform in Texas. Retrieved from https://www.texascjc.org/system/files/publications/Reversing%20the%20Pipeline%20Report%202020.