Nick Moir & Matthew Gottfried

DS 2002

Professor Jason Williams

Final Project Data Report

DATA SELECTION

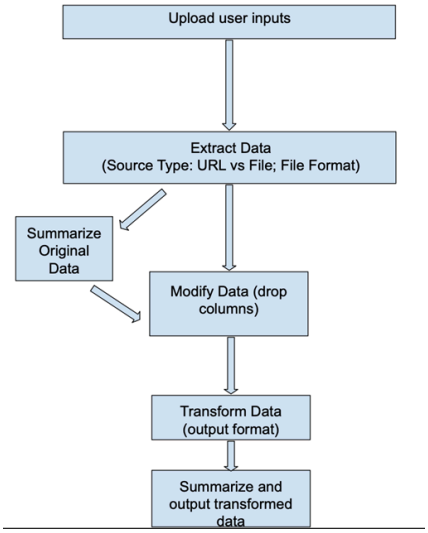
When we were tasked with choosing data sources for this project, we assumed that it would be easy to find datasets online that perfectly matched our needs. After hours of trying to find promising sources, we realized it wasn't as simple as we thought. One thing we noticed immediately was how hard it was to find data for matching geographics. When trying to use data to analyze societal trends, the geography of the data matters a lot. Some datasets provided data for each country, some for each state, and others for towns or counties. We had to find two different datasets that broke down the geography of their data in precisely the same way to adequately draw real-world analysis from the data.

To address this issue, we decided to specify our search to be only datasets focused on the counties of Virginia. This allowed for very localized data, so patterns could be observed in some of the smallest population places in Virginia without their insights being lost in Virginia as a whole. Once we had narrowed our search down to the counties and cities of Virginia, we had to find a trend to analyze. We took some time to think about matters we have observed and wanted to quantify. Nick mentioned that he thought it was interesting that whenever UVA students would receive a text about a shooting or burglary in the area, they tended to be primarily in low-income neighborhoods. It seemed clear that there was a connection between levels of income in a community and the crime rate, and we thought it was something that could be explored well through data.

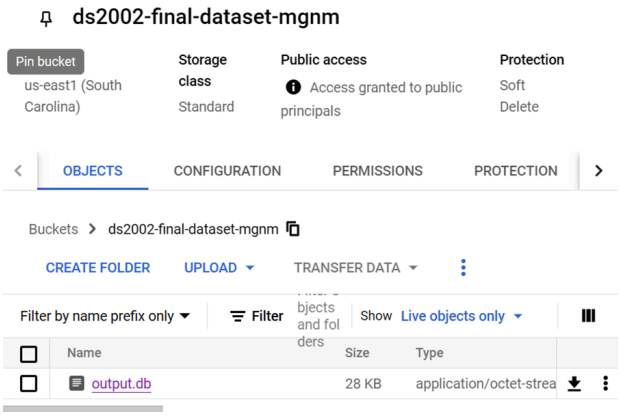
After identifying our focus, we found that the National Institute of Health turned out to be the best source for reliable data on Virginia cities and counties. We chose to measure median household income and the rate of people without a high school diploma, as education can affect income, and in turn, violence. All the data covered 2018-2022 and provided us with the data we need for our independent variables for regression models. Next, we found incarceration data from the nonprofit Prison Policy Initiative. This organization uses research and advocacy to try to end mass incarceration. This data perfectly matched the geography and timeline of our National Institute of Health Data.

DATA TRANSFORMATION, CLEANING AND CLOUD STORAGE

After collecting our data, we were left with multiple CSV files that were over-cluttered with unnecessary columns. To solve this issue, we used our ETL pipeline to combine the files into one while simultaneously dropping all of the columns that we deemed unnecessary. This step was critical to our success; if we had to do all of this by hand, it would have taken significantly longer. When creating our ETL pipeline, we looked at many different steps in the process. The first hurdle we had to move past was extraction. We wanted to ensure that our pipeline could handle a data source from a file or a URL. After completing the extraction, we moved on to transformation. In our pipeline, there is some level of user interaction, in which you can specify what columns you would like to drop. After completing your data cleaning, the user selects the type of file they want the new file to be. The pipeline then converts the file to the specified type and saves it to the user's working directory. After developing our pipeline, we used it to convert our CSV file into a SQL file. We opted to use SQL because of our experience using this language.



When considering cloud storage, we used Google Cloud. The first step was creating a bucket. Because there is only one file, we did not have to worry about organizing the bucket. To manage access, we created a service account to gain access to the bucket. We also assigned ourselves as role/storage.admin. To ensure that no unauthorized changes have been made, we can review the observability tab. We used BigQuery to ensure that we were able to access the data.

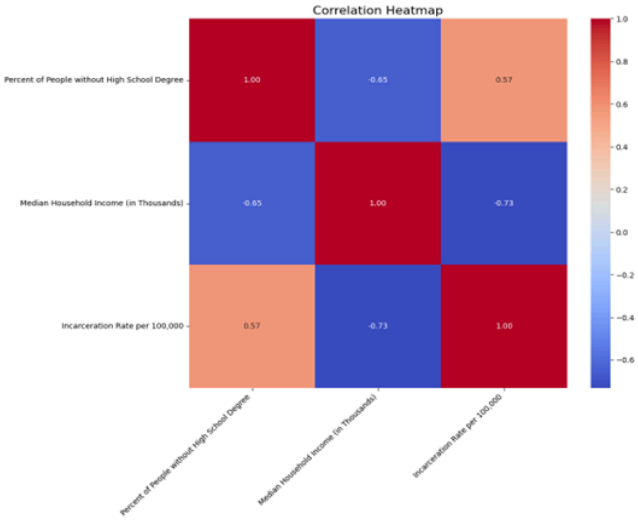


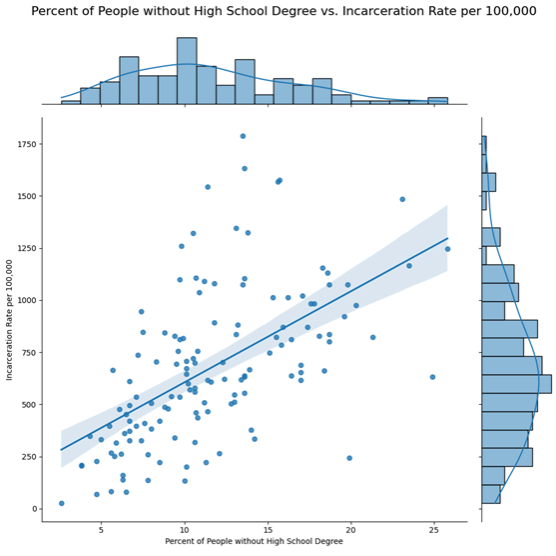
DATA ANALYSIS

Once the data was all cleaned, transformed, and stored in a single database file, we could begin analyzing the data for patterns. We used “percent of people without a high school degree” and “median household income” as our independent variables, as neither would be very affected by any small disparities in the general population from any other dataset. We also chose “incarceration rate per 100,000” to serve as our dependent variable, for the same reason: to minimize bad data from population disparities. Once we had our variables, we first performed a simple summary statistics calculation on each of the three metrics. These included total row count, mean, standard deviation, and each quartile value. This data was not as useful to us, as it did not analyze any connection between the variables, just the makeup of each column. The one surprising fact was that the standard deviation was the highest relative to its mean for incarceration rate, meaning it varies the most from town to town. We both had expected income to be the largest relative standard deviation.

Next, we wrote code for a Pearson correlation matrix using the Scipy.Stats package to analyze the correlation coefficient between each variable and visualize it on a graph. To do this, we used the heatmap

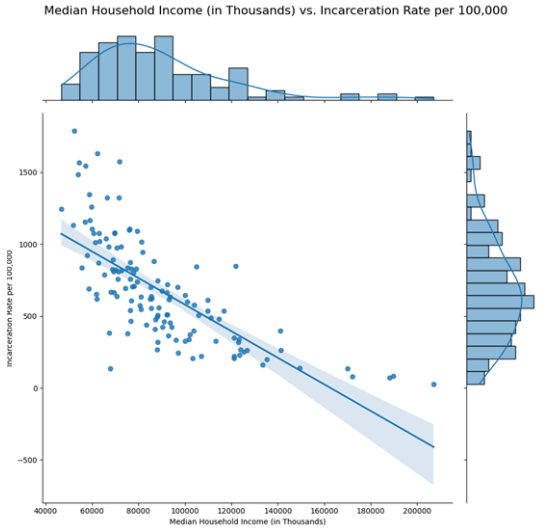
function in the Seaborn package to generate an image with the 3x3 matrix we generated. It allows viewers to visualize how strong the correlation is between income and incarceration rates, further helped by the use of blue and red to signify intensity. The Pearson correlation coefficient for the relationship between median household income and incarceration rate was -0.73, which indicates a strong correlation between variables. This is followed by a correlation coefficient of 0.57 for the percent of people without a high school degree and incarceration data, which is a more moderate relationship than the former. These indicate a solid correlation between both variables and the incarceration rate, but it is clear that median household income has a greater effect on the community's incarceration rate than the percentage of people without a high school degree.





Finally, we used linear regression to measure the performance of the regression between both independent variables and the dependent variable. Using various packages from Sklearn, we measured the mean squared error to find the mean squared difference between predicted and actual values, essentially measuring spread. We also measured the R^2 score for variance in the dependent variable explained by variance in the independent variable. Finally, we calculated the coefficient of the line of best fit, the Pearson correlation, and the p-value to

measure statistical significance. We plotted the line of best fit on a scatter plot using Matplotlib to show the spread and included two other tables to show the distribution of each variable. When analyzing both the numerical analyses, we reached a split conclusion. The prediction for the data in median household income vs. incarceration rate is far less accurate than the prediction for the data in the percent of people with less than a high school diploma, with a lower mean squared error (32,817 vs. 53,108) and a higher R^2 value (0.587 vs. 0.322). However, the correlation between median household income and the incarceration rate is clearly stronger, as indicated visually and by the higher Pearson correlation (-0.731 vs. 0.572) and a smaller p-value (1.81e-23 vs. 6.47e-13). After further visual analysis, it seems that the relationship may not be linear, which would cause the linear regression to be a worse-performing metric.



	Percent of People without High School Degree	Median Household Income (in Thousands)	Incarceration Rate per 100,000
count	133.000000	133.000000	133.000000
mean	11.643609	89408.857143	677.706767
std	4.804549	28989.533481	366.777816
min	2.600000	46783.000000	27.000000
25%	7.800000	69671.000000	396.000000
50%	10.700000	83470.000000	637.000000
75%	14.200000	101797.000000	872.000000
max	25.800000	207090.000000	1787.000000

```
Linear Regression for Percent of People without High School Degree vs. Incarceration Rate per 100,000:
  Mean Squared Error: 32817.76133986597
  R^2 Score: 0.5871525509727759
  Coefficient: 43.36051216185541
  Pearson Correlation: 0.5718959856213365
  P-Value: 6.470183356475604e-13

Linear Regression for Median Household Income (in Thousands) vs. Incarceration Rate per 100,000:
  Mean Squared Error: 53108.21396259927
  R^2 Score: 0.33189864994792895
  Coefficient: -0.009276351522551362
  Pearson Correlation: -0.7307134990152957
  P-Value: 1.8143763855607313e-23
```