Incarceration Correlation: The Roles Race and Region Play on Incarceration Rates in the United States

Ashley Jimenez, Brandon Milligan, Jeremiah Reynoso and Kamryn Robertson
Advised by: Dr. Kendra E. Pleasant

July 2021

Abstract

The 2010 U.S. Census was one of the few data sources that contained state-level data that allowed an analysis, by gender or race/ethnicity, of the racial disparities in the criminal justice system. The Census was broken into counts of incarcerated people by race for the total population, for men, for women, and for the combined population of adult men and women. In this study, we aim to find significant correlation in the incarceration rates of the African American, Caucasian, Hispanic, and Native Hawaiian and other Pacific Islander races/ethnicities by applying Ramsey's Theory to complete graphs.

Contents

1 **Mathematical Background** **3**

1.1 Graph Theory . . . 3

1.2 Ramsey Theory . . . 5

2 **Data Application** **8**

2.0.1 African American . . . 9

2.0.2 Caucasian . . . 10

2.0.3 Hispanic . . . 10

2.0.4 Native Hawaiian/Pacific Islander . . . 10

2.1 Northeast . . . 11

2.2 Midwest . . . 11

2.3 West . . . 12

2.4 South . . . 13

3 **Future Works** **14**

1 Mathematical Background

This project is an application of Ramsey theory. In this section of the paper, we will be describe the mathematics needed to achieve our goal.

1.1 Graph Theory

Typically, when we think about a graph we think about the graph of a function. Suppose f is a function mapping a finite set A to itself. Then the graph G of f is defined as

$$G = \{(x, f(x)) : x \in A\}.$$

Let's change the way we are defining this graph. Suppose we let each element in A be represented by a vertex or node. If the coordinate point (x, y) is in G then a line or edge will be placed on the graph between the vertices labeled x and y. This shows that x and y are paired together. This defines a new visualization of the graph G. A formal definition is given below.

Definition 1. *A* ***graph*** *consists of two finite sets, V and E. Each element of V is called a* ***vertex****. The vertex set of a graph G is denoted by $V(G)$ or simply V . The elements of E are unordered pairs of vertices called* ***edges****. An edge connecting vertices u and v is denoted uv and u and v are said to be the* ***edge's ends****. The edge set is denoted by $E(G)$ or simply E. The graph is denoted $G = (V, E)$.*

This new definition allows us to visualize graphs outside of the typical $xy-$coordinate system. Now, we just need a collection of vertices, $V = V(G)$, and a collection of edges, $E = E(G)$. Let's analyze our new graphs more deeply.

Definition 2. *Let $G = (V, E)$ be a graph where $V = V(G)$ and $E = E(G)$ are finite sets. Then*

1. *the* ***order*** *of a graph G is the cardinality of $V(G)$, denoted $|V(G)|$, and*
2. *the* ***size*** *of a graph G is the cardinality of $E(G)$, denoted $|E(G)|$.*
3. *Two vertices u and v in $V(G)$ are said to be* ***adjacent*** *if uv is one of the edges in $E(G)$.*

Essentially, we can say that the order of a graph G is the number of vertices on the graph and the size of a graph G is the number of edges on the graph. We will solidify these ideas with the following example.

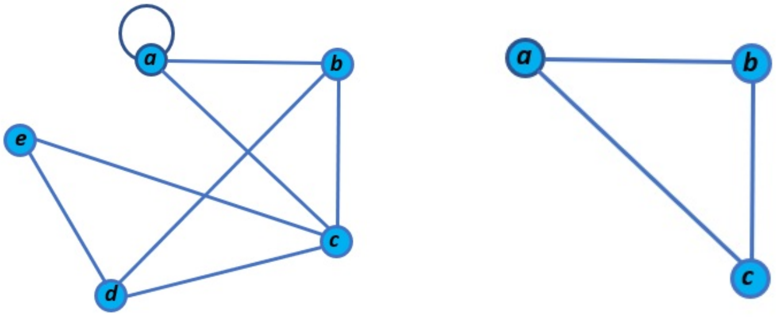
Example 1. *Let G and H be graphs where $G = (V(G), E(G))$ and $H = (V(H), (E(H))$. Suppose*

- *$V(G) = \{a, b, c, d, e\}$, $E(G) = \{ab, aa, cd, ec, bc, db, ac, de, bd\}$,*
- *$V(H) = \{a, b, c\}$, and $E(H) = \{ab, ac, bc\}$.*

Let's graph G and H. First, let's graph G. We start by drawing each vertex in the set $V(G)$. If an edge uv is in $E(G)$, then we draw a line connecting the vertices labeled u and v. For instance, the edge ab is in $E(G)$. So the graph G needs a line connecting the vertices labeled a and b. Since the edges are not ordered then the edge ab can also be represented as ba. Now let's graph H. Similarly as before, we start by drawing the vertices from the set $V(H)$. Then we connect our vertices with the edges from the set $E(H)$. The visual representation of G and H can be found in Figure 1.

Notice, there is a edge connecting a to a in the graph of G. An edge whose edge end's are the same vertex is called a **loop**. The loop makes a circle around the vertex labeled a. Also notice that there are two edges whose edge end's are b and d. This is called a **multiple edge**. When a graph contains a loop or a multiple edge it is called a **multigraph**. For the remainder of this paper, we will focus on *simple graphs*. Let's formally define a simple graph.

Definition 3. *A graph is said to be* ***simple*** *if it does not have any loops of multiple edges.*



(a) G (b) H

Figure 1: Visual representations of graphs G and H.

Consider graphs G and H from the previous example. We notice that there is a relationship between the graphs' sets of vertices and edges respectively. We notice that $V(H)$ is a subset of $V(G)$ and $E(H)$ is a subset of $E(G)$. When this is true, we call H a *subgraph* of G. Let's formally define a subgraph.

Definition 4. *Let G and H be two graphs. Then H is said to be a* ***subgraph*** *of G if $V(H) \subseteq V(G)$ and $E(H) \subseteq E(G)$.*

Notice that all of the vertices in $V(H)$ are adjacent to each other. This implies that H is a special type of subgraph. To fully understand this special type, we must introduce a *complete graph.*

Definition 5. *Let V be a set of n vertices. Then graph G is called the* ***complete graph*** *on V if for all u and v in V there exists edge uv in $E(G)$. The complete graph is denoted K_n.*

For each positive integer n the complete graph K_n is unique. The following theorem allows us to find the size of the complete graph created using n vertices. This will allow us to analyze the complete graph without having to physically draw it.

Theorem 1. *Let $V = \{v_1, v_2, v_3, \ldots, v_n\}$. Then the complete graph K_n has $\dfrac{n(n-1)}{2}$ edges.*

Proof. We will show $|E(K_n)| = \frac{n(n-1)}{2}$. For each $i, j \in \{1, 2, 3, 4, \ldots, n\}$ then element v_iv_j must appear in $E(K_n)$. Note that $i \neq j$ and $v_iv_j = v_jv_i$. As a result, order does not matter and each item i can only be used once in each v_iv_j pairing. This implies that we can use a combination to find $|E(K_n)|$. We want to choose any two vertices from a collection of n vertices. Therefore, $|E(K_n)| = \dbinom{n}{2}$ where

$$\begin{aligned}\binom{n}{2} &= \frac{n!}{2!(n-2)!}\\ &= \frac{n(n-1)(n-2)!}{2!(n-2)!}\\ &= \frac{n(n-1)}{2}.\end{aligned}$$

□

Recall the graphs G and H from Example 1. We mentioned earlier that H is a special type of subgraph. We now can say that H is a complete graph created using three vertices. So, we call H a *triangle subgraph.*

Definition 6. *Let H and G be graphs where H is a subgraph of G. If H is itself a complete graph with order three then we say H is a* ***triangle subgraph****.*

In section two of this paper, we will focus on a complete graph and its triangle subgraphs. Thanks to Definition 1 we know that a graph can be represented by a collection of vertices and a collection of edges. The following definition will allow us to visualize a graph as a finite matrix.

Definition 7. *Let n be a positive integer. A $n \times n$ matrix $A(G) = (a_{i,j})$ is called an* ***adjacency matrix*** *if*

1. *$a_{i,j} = 0$ or $ai, j = 1$ for each i and j,*
2. *$a_{i,i} = 0$ for each i, and*
3. *$A(G)^T = A(G)$.*

Example 2. *Consider the graph K_6. Recall that K_6 is the graph with six vertices where any two vertices are adjacent to each other.*

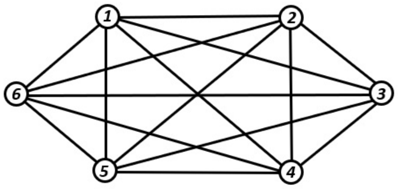


Figure 2: Visual representation of the complete graph K_6.

We want to build the adjacency matrix, $A(K_6)$, for K_6. Since K_6 has six vertices then $A(K_6)$ is a 6×6 matrix and we can write $V(K_6) = \{1, 2, 3, 4, 5, 6\}$. By definition, we place a one in the matrix position $a_{i,j}$ if there is an edge connecting the vertices labeled i and j, and we place a zero in the matrix position $a_{i,j}$ if there is not an edge connecting the vertices labeled i and j. Since the complete graph has an edge between any two vertices then $a_{i,j} = 0$ if and only if $i = j$. Hence

$$A(K_6) = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}.$$

We recognize that there is a one-to-one correspondence between a graph G and its adjacency matrix $A(G)$. For this reason, the two ideas will become synonymous for the rest of this paper. This means that we can treat an adjacency matrix as the graph it represents and vice versa. Now that we have a firm understanding of complete graphs, we want to begin placing finite colorings on them.

1.2 Ramsey Theory

The first thing we need to discuss is a *finite coloring.*

Definition 8. *Let k be a natural number and A be a nonempty set. A function $f : A \to \{1, 2, 3, ..., k\}$ is called a* ***finite coloring*** *of A. If such a coloring exists we say that A is* ***finitely colored****. For each $i = 1, 2, 3, \ldots, k$, the set $C_i = \{a \in A : f(a) = i\}$ is called a* ***color class of*** $\mathbf{f}$.

Since f has k outputs then we say that f is a $k-$coloring of A. There are two types of $k-$colorings that can be placed on a graph $G = (V, E)$. We can either color the set of vertices V or the set of edges E.

Definition 9. *Let k be a natural number and $G = (V, E)$ be a graph.*

1. *A* **k**$-$***coloring of the vertices*** *is a function $f : V \to \{1, 2, 3, ..., k\}$ such that if $x, y \in V$ where x and y are adjacent to each other then $f(x) \neq f(y)$. If such a coloring exists we say V is* **k**$-$***colorable***.
2. **k**$-$***coloring of the edges*** *is a function $f : E \to \{1, 2, 3, , ..., k\}$.*

In this paper, we will be placing 2$-$colorings on the edges of complete graphs. The first color class will commonly be referred to as the color red, and the second color class will be commonly referred to as the color blue. As mentioned earlier, we care about finding triangle subgraphs. In fact, we want the triangle subgraphs to belong to one color class.

Definition 10. *Let A be a set that has been finitely colored. That is, there exists $f : A \to \{1, 2, 3, \dots, k\}$ for some k. We say that a subset B of A is* ***monochromatic*** *if $f[B] \subseteq C_i$ for some $1 \leq i \leq k$.*

We mentioned earlier that we will place a 2$-$coloring on the edges of a graph and find triangle subgraphs that belong to one color class. So a *monochromatic subgraph* will refer to a subgraph whose edges belong to one color class. We will now provide a concrete example of a finite coloring of a graph.

Example 3. *Consider K_6 from Example 2.*

We want to place a finite coloring on the nodes and a finite coloring on the edges. By definition, to place a coloring on our vertices we must have each node in a different color class from its adjacent vertices. Since we are considering the complete graph, then all of the vertices are adjacent with each other. As a result, each vertex must be assigned a different color from the rest. We also note that there is no restriction on how the edges are colored. A finitely coloring of K_6 is as follows. Let $f_1 : E(K_6) \to \{1, 2\}$ where

- $f_1[\{12, 23, 34, 45, 56, 16, 14, 25, 36\}] = \{1\}$ *and*
- $f_1[\{13, 24, 35, 46, 26, 15\}] = \{2\}$.

Let $f_2 : V(K_6) \to \{1, 2, 3, 4, 5, 6\}$ where $f_2(i) = i$ for each $i = 1, 2, 3, 4, 5, 6$. A visual representation of this coloring can be found below.

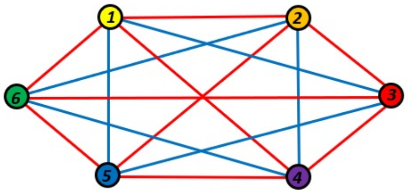


Figure 3: Visual representation of the 2$-$coloring on the edges and vertices of K_6.

Let $H = (V(G), E(G))$ where $V(G) = \{1, 4, 5, 6\}$ and $E(G) = \{12, 25, 56, 16\}$. Then H is a subgraph of G whose edges are monochromatic.

We saw in the previous section that we can use an adjacency matrix to represent a graph G. We can also use finite matrices to represent a 2$-$coloring on graph G.

Definition 11. *Let G be a 2-colored graph with adjacency matrix $A(G)$. Define $R(G) = (r_{i,j})$ and $B(G) = (b_{i,j})$ to be two $n \times n$ matrices where*

$$r_{i,j} = \begin{cases} 1, & a_{i,j} \text{ is a red edge} \\ 0, & a_{i,j} \text{ is a blue edge} \end{cases} \text{ and } b_{i,j} = \begin{cases} 1, & a_{i,j} \text{ is a blue edge} \\ 0, & a_{i,j} \text{ is a red edge} \end{cases}.$$

Note that $A(G) = R(G) + B(G)$. The matrix $R = R(G)$ will be used to represent the collection of red edges and $B = B(G)$ will be used to represent the collection of blue edges. We can use these matrices to count the number of monochromatic triangles.

Theorem 2. *Let A be an $n \times n$ adjacency matrix for a $2-$colored graph $G = (V, E)$, and let R and B represent the red and blue edges of G respectively. Then the number of monochromatic triangles M in G is given by*

$$M = \frac{Trac(R^3) + Trac(B^3)}{6}.$$

Proof. [3, Corollary 2] □

In Section 2 of this paper, we will turn a data set into a complete graph. We will then use Ramsey theory to find correlation in said data set. The rest of Section 1.2, will set the stage for the analytics that will be done in Section 2.

Theorem 1. *Let p and q be integers. There exists a smallest positive integer n such that every $2-$coloring of K_n has a complete subgraph of p vertices whose edges are all colored red or a complete subgraph of q vertices whose edges are all colored blue.*

Proof. [2, Theorem 5.1] □

The smallest positive integer n is called the **Ramsey number** and is denoted $R(p, q)$. As mentioned earlier, we want to find the existence of monochromatic triangle subgraphs. In other words, we need $p = q = 3$.

Theorem 2. $R(3, 3) = 6$

Proof. [2, Theorem 5.4] □

This tells us that as long as a complete graph has more than six vertices than any $2-$coloring of it will have a red triangle subgraph or a blue triangle subgraph. Suppose G is a $2-$colored complete graph. We want to determine what percentage of the triangle subgraphs of G are monochromatic. First we will find the total amount of triangle subgraphs on graph G.

Theorem 3. *Let K_n be a complete graph. Then there are $\binom{n}{3}$ triangle subgraphs of K_n.*

Proof. Let $V = \{v_1, v_2, \ldots, v_n\}$ be the vertex set for K_n. Recall that a triangle subgraph is a complete graph with three vertices. Let $W = \{v_i, v_j, v_k\}$ be a subset of V. Since K_n has an edge connecting any two vertices in V then we know there is a triangle subgraph of K_n whose vertex set is W. We need to find the total number of ways W can be built. This implies that we can use a combination to find the number of triangle subgraphs of K_n. We want to choose any three vertices from a collection of n vertices. Therefore, the number of triangle subgraphs of K_n is equal to $\binom{n}{3}$ where

$$\begin{aligned}\binom{n}{3} &= \frac{n!}{3!(n-3)!} \\ &= \frac{n(n-1)(n-2)(n-3)!}{3!(n-3)!} \\ &= \frac{n(n-1)(n-2)}{6}.\end{aligned}$$

□

We will be using this count to find the percentage of monochromatic triangle subgraphs on a specific complete graph that has been finitely colored. In 1959, A. W. Goodman was able to find a formula for the minimum amount of monochromatic triangles that are guaranteed for a complete graph with n vertices. This result is given in the following theorem.

Theorem 3. *Let G be a graph with n vertices and edges colored either red of blue. Then the quantity of monochromatic triangles, T_n, in G is at least*

$$T_n = \begin{cases} \frac{m(m-1)(m-2)}{3}, & n = 2m \\ \frac{2m(m-1)(4m+1)}{3}, & n = 4m+1 \\ \frac{2m(m+1)(4m-1)}{3}, & n = 4m+3 \end{cases}.$$

Proof. [1, Theorem 1] □

From this formula, Goodman was able to create an estimate for the percentage of monochromatic triangles an arbitrarily $2-$colored complete graph K_n should have. He was also able to show that this estimate approaches twenty-five percent as n approaches infinity.

Theorem 4. *Let G be a $2-$colored graph with n vertices. Then the percentage of triangle subgraphs of G that are monochromatic is asymptotically at least* $\frac{n-3}{4n} \to \frac{1}{4}$.

Proof. [3, Corollary 3] □

The estimated percentage, $\frac{n-3}{4n}$, will commonly be referred to as $Goodman(n)$. To finalize the idea of finding meaningful connections in our data sets, we must also define $Mono(G)$. We will compare $Mono(G)$ to $Goodman(n)$.

Definition 12. *Let G be a $2-$colored graph with n vertices. Let M be the number of monochromatic triangle subgraphs of G given by Theorem 2. Then the percentage of monochromatic triangle subgraphs of G, denoted $Mono(G)$, is given by:*

$$Mono(G) = \frac{M}{\binom{n}{3}}.$$

Suppose G is a graph with n vertices whose edges are colored blue or red. Let $Mono(G)$ be the percentage of monochromatic triangle subgraphs of graph G. If

$$Mono(G) > Goodman(n)$$

then we say G has potentially meaningful connections. We will now use these ideas to determine if a data set of incarceration rates has an meaningful connections.

2 Data Application

This opening paragraph is not useful. We did not curate this data from various websites. You should be referencing the one website you got the data set from. You may say that they curated the data that way. It seems as if you are taking the credit.

The data we used for our research came from the Prison Policy Initiative. They compiled their data using information on the 2010 census regarding the total number of people incarcerated throughout the United States. Their data also compiled the total number of incarcerated folks per $100,000$. They divided the total prison population of a certain race (per state) by the total population of that particular race. Finally, they multiplied the quotient by $100,000$. We used the incarceration per $100,000$ for all of our calculations.

As mentioned earlier, we want to determine if there is any correlation between the incarceration rates across this country. The data set we are using includes all fifty states and the District of Columbia. We aim to determine if there is any correlation between the incarceration rates of each race. We will highlight the following races: Caucasian; Black/African American; Native Hawaiian/Pacific Islander; and Hispanic. We had a two research questions in mind at the beginning of this project.

1. Is there a connection between race and incarceration rates in the United States?
2. Are there differences between races with regards to incarceration, and is this difference nationwide or state specific?

We will view this data set as a $2-$colored complete graph. In order to define the $2-$coloring, we need to introduce a *Hamming distances* and a *Hamming matrix*.

Definition 13. *Let S be a data set whose elements are vectors of length n. Let $\vec{x} = (x_1, \dots x_n)$ and $\vec{y} = (y_1, \dots, y_n)$ be two elements in S. Then the* ***Hamming distance*** *from $\vec{x}$ to $\vec{y}$, denoted $h_{x,y}$ is given $h_{x,y} = |i : x_i \neq y_i|$*

Note that our data set is a list of incarceration rates as opposed to a collection of vectors. As a result, we will be using a slightly modified Hamming distance.

Definition 14. *Let S be a subset of $\mathbb{R}$. Let x and y be two elements in S. Then the **Hamming distance** from x to y, denoted $h_{x,y}$ is given by*

$$h_{x,y} = |x - y|.$$

We will use the Hamming distances to build a finite matrix H.

Definition 15. *Let S be a finite data set. Pick positive integer k such that S has k elements and write $S = \{s_1, s_2, s_3, \dots, s_k\}$. The **Hamming matrix**, denoted H, is a $k \times k$ matrix where each element $h_{i,j}$ is the Hamming distance from s_i to s_j.*

Now we can discuss our process. Please be aware that this process was completed using `R Studio`. We will be analyzing the incarceration rates for four races. The first step in our process was to createIn our data application, we are taking values and comparing it with other elements within that particular data set, hence why we are focusing on Hamming matrices and creating complete graphs from that data set. As a result, our graph will have the same number of elements as were present in the hamming matrix. For instance, if our hamming matrix had 4 individual elements, then we could expect our graph to have 4 vertices, and vice versa.

It is important to mention that complete graphs do not show values within the edges, which is why we utilize a 2-coloring to clearly distinguish between various values. We decided that the most effective strategy in parsing through our huge data set was to create 'Thresholds' within the graph. But before we created our graphs, we found the max hamming distance (The max value of the set subtracted by the lowest). Let us note that each race will have a different maximum hamming distance. So, when we find our maximum hamming distance value, we need to divide that number by 20 and add our new value to the previous value for 21 thresholds. For example, if the max hamming distance for a particular set is 1,000, then we would create thresholds of 50 (including zero) - these increments would now be used on the graph to determine whether an edge should get a red or blue coloring. All of our 2-coloring complete graphs follow this methodology. In our graphs, if the value threshold is greater or equal to the hamming distance of a particular edge, then that edge would get the blue coloring - values lower than the threshold would get a red coloring.

The importance of this coloring is that we want to find the maximum number of monochromatic triangles each threshold creates, then dividing that value by the total number of triangles that are seen on the complete graph, whereby, giving us Mono(g). You would then compare this value with Goodman's number. If our Mono(g) > Goodman(n), then we can say that there's a significant correlation in the data set.

We have recorded our significant findings below:

2.0.1 African American

When we considered all fifty states for the African American race it was found that the max hamming distance was 7360.117. With this number it was decided to place increments of 369 on our threshold. After analyzing each threshold, we found that for every threshold t, Mono(g) was greater than Goodman(n). By looking at our original data set, we found that the max Hamming distance of 7360.117 came from the comparison of the incarceration rates of the Southern regions West Virginia and the District of Columbia. This means that there is less consistency in the incarceration rates of the South region.

In contrast, when we look at the minimum hamming distance, 0.501, it was decided to place increments of 52 on our threshold. After analyzing each threshold, we found that for every threshold t, Mono(g) was greater than Goodman(n). By looking at our original data set, we found that the min Hamming distance of 0.501 came from the comparison of the incarceration rates of the Southern regions Hawaii and the District of Columbia. This shows us some consistency in our data.

2.0.2 Caucasian

When we considered all fifty states for the Caucasian race it was found that the max hamming distance was 799.5435. With this number it was decided to place increments of 40 on our threshold. After analyzing each threshold, we found that for every threshold t, Mono(g) was greater than Goodman(n). By looking back at our original data set, we found that the max Hamming distance of 799.5435 came from the comparison of the incarceration rates of the South region states District of Columbia and Oklahoma. This means that there is less consistency in the incarceration rates of the South region.

Now when considering the smallest max hamming distance, it was 171.7289. With this number it was decided to place increments of 8.6 on our threshold. After analyzing each threshold, we found that for every threshold t, Mono(g) was greater than Goodman(n). By looking back at our original data set, we found that the max hamming distance of 171.7289 came from the comparison of the incarceration rates of the Northeast region states Vermont and Pennsylvania. This means that there is more consistency in the incarceration rates of the northeast region.

We notice that the Caucasian population is missing the plateau that the other races have in their graphs. We recognize that the plateau in the graphs means there is no significant change in the way the edges are colored in our graphs. This means the Caucasian thresholds are more closely related, more consistent, and each threshold gives us a constant change in our triangle colors

2.0.3 Hispanic

When we considered all fifty states for the Hispanic race, we found that the max hamming distance was 4618.691. With this number it was decided to place increments of 231 on our threshold. After analysing each threshold, we found that for every threshold t, Mono(g) was greater than Goodman(n).After initially analyzing the data of all 50 states, you recognize that the states that created our max hamming distance were Delaware and Mississippi - Mississippi being deep in the South and Delaware in the Northeast. From this preliminary information, we can already see the trends that are going to be seen throughout our data - huge disparities with regards to the South versus other regions.

Comparatively, when we look at the lowest hamming distance, we get 6.2665 - coming from New York and Washington. And while these numbers don't signify that New York and Washington have the lowest incarceration numbers, they do help us notice the consistency that is seen in our data.

We notice that when we progress through our thresholds, we see various sections that 'plateau' in the data set. As a result of our data not being perfectly balanced throughout, you see these gaps to which nothing significant changes with respect to how many monochromatic triangles are present on the graph.

2.0.4 Native Hawaiian/Pacific Islander

When we considered all fifty states for the Native Hawaiian/ Pacific Islander races it was found that the max Hamming distance was 8135.674. With this number it was decided to place increments of 420 on our thresholds. After analysing each threshold, we found that for every threshold t, Mono(G) was greater than Goodman(n). By definition, this means that there was meaningful correlation for each analyzed value of t.

By looking back at our original data set, we found that the max Hamming distance of 8135.674 came from the comparison of the incarceration rates of the western region states Arizona and Montana. In order to figure out why these two states were significant in our data analysis, our data set was broken into four regions. It was then found that Arizona significantly caused a difference in the percentage of monochromatic triangle subgraphs from our other western region states.

We notice that not only in every analysis of our Mono(G) percentages that we reach areas where the percentages do not change, or what we call a plateau, but we also begin to see a decrease in the number of blue monochromatic triangles until B=0 for some threshold t. However, the number of monochromatic blue triangles in the western region dies out fairly quicker than any other race. With further investigation, it was found that by removing Arizona, our Hamming differences decreased. This implies that Arizona was in fact significantly impacting the max Hamming distance when all fifty states were considered for the Native Hawaiian/ Pacific Islander races.

In each subsubsection of the paper you each need to explain your findings when considering all fifty states. Be sure to mention the max Hamming distance and how many of the thresholds that had significant correlations. Be sure to include why we decided to break the races up by region.

2.1 Northeast

We recognize the following states are apart of the Northeast region: Connecticut, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, and Vermont.

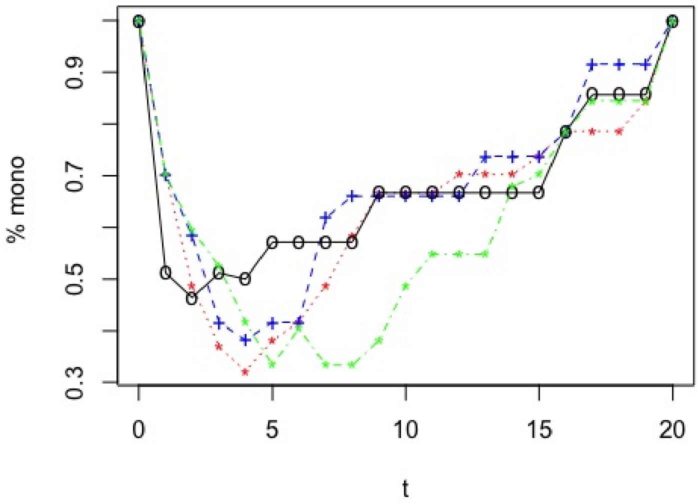


Figure 4: Visual representation of Northeast Region Correlation.

We notice that if the hamming distance is smaller than our threshold value t, then the edges of our monochromatic triangles will be colored red and if the hamming distance is greater than or equal to our threshold t, then the edges our monochromatic triangles will be colored blue.

We notice that the number of blue monochromatic triangles for African American, Hispanic, and Natives all hit 0 at the same threshold number indicating that the max hamming distances for these races in the northeast had to be relatively close as the number of blue monochromatic triangles becomes 0 at threshold 10.

2.2 Midwest

We recognize the following states are apart of the Midwest region: Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, and Wisconsin.

When we looked at the max hamming distances for the different races, we noticed the same states come up: South Dakota, Wisconsin, and Nebraska. These states being responsible for having either the most or the least of a certain race incarcerated. South Dakota had the highest number of incarcerated African American and Hispanic, while Wisconsin had the most Native Hawaiian incarcerated.

When looking at the various thresholds for the Midwest region, we saw a pattern begin to take shape with regards how many Monochromatic triangles appeared on the graph. We called these increments 'plateaus' because that's how they appear on the graph - flat horizontal lines. The reason for these plateaus is because our data set is not balanced. We have these outliers both as a maximum and a minimum, as a result, we have a range of thresholds that don't significantly change the number of monochromatic triangles because a bulk of our values are either above that specific threshold or below it. We also found it interesting that the plateaus happen around the same threshold count - again, speaks to how big the range is of our data set.

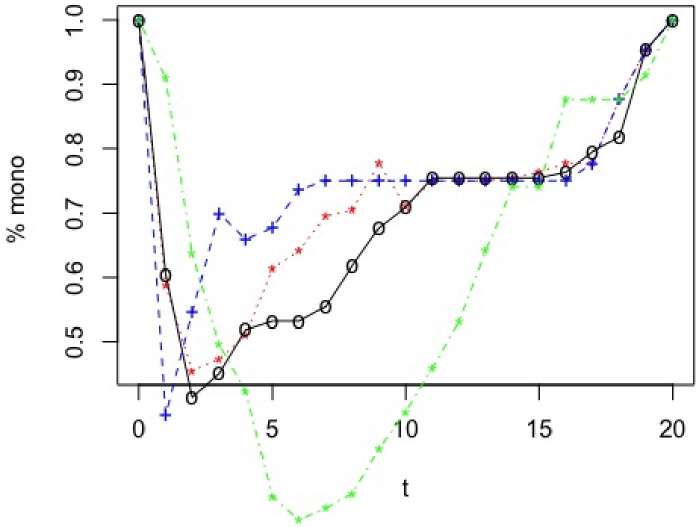


Figure 5: Visual representation of West Region Correlation.

2.3 West

We recognize that the following states are apart of the West region:
Alaska, Arizona, California, Colorado, Hawaii, Idaho, Montana, Nevada, New Mexico, Oregon, Utah, Washington, and Wyoming.

We note that if the hamming distance is smaller than our threshold value, t, then the edges of our monochromatic triangles will be colored red and if the hamming distance is greater than or equal to our threshold value, t, then the edges of our monochromatic triangles will be colored blue.

In our data analysis of the maximum hamming distance, change in threshold for each race, and of the states in the west region, we find a significant relationship between this coloring of the edges. We first notice that when the number of our blue monochromatic triangles begins to reach 0, we can interpret this as the majority of distances becoming smaller than our threshold values. For instance, we note that the Native Hawaiian/ Pacific Islander races significantly differs in rate at which this begins to happen from the other races. We also see that our majority of differences between incarceration rates for the Hawaiian/ Pacific Islander races is significantly larger than all other races as well.

By comparing these values of hamming distances, threshold values, individual incarceration rates of each state in the west region, we can conclude that there is more of a variety in our incarceration rate differences in our higher numbers, which in this case specifically came from the population of Native Hawaiian/ Pacific Islanders in the state of Arizona.

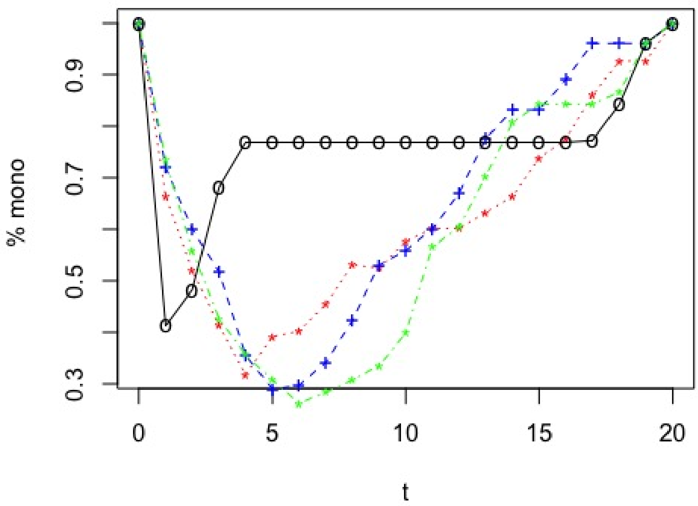


Figure 6: Visual representation of West Region Correlation.

2.4 South

We recognize the following states are apart of the South region: Alabama, Arkansas, Delaware, District of Columbia, Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia, and West Virginia.

When we are looking at the max hamming distances for the different races, we see that Oklahoma had the most Caucasian people incarcerated (889), West Virginia had the most African American people incarcerated (7360) and Native Hawaiian people incarcerated (3271), and Mississippi had the most Hispanic people incarcerated (4869).

When looking at the various thresholds for the South region, we saw a pattern take shape with regards to how many monochromatic triangles appeared on the graph. We are noticing plateaus, flat horizontal lines, because our data set is not balanced. Our plateaus are happening mainly with our Hispanic and Native Hawaiian/Pacific Islander races.In our data analysis, we notice that the number of blue monochromatic triangles for African American, Hispanic, and Native Hawaiian/Pacific Islander all hit 0 at the same threshold number, which is 7. This indicates that the max hamming distances for these races in the south had to be relatively close as the number of blue monochromatic triangles becomes 0 at threshold 7. Moreover, at threshold 16, the blue monochromatic triangles for Caucasians hit 0.

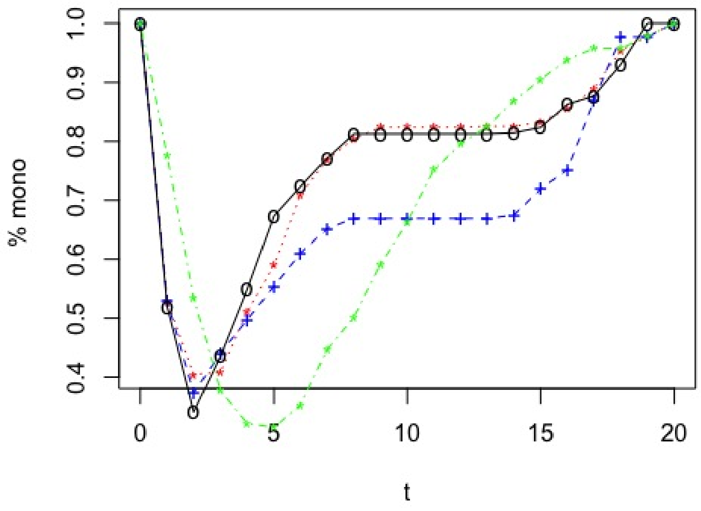


Figure 7: Visual Representation of South Region Correlation.

In each of the previous section you all need to include what states are being included in each region. You also should include the plots created by Brandon and Jeremiah. You can then discuss the interesting things that are happening such as the plateus, the fact that B will eventually equal zero and what that implies about the big picture, and the states with the highest incarceration rates for each region.

3 Future Works

When looking into the data, we looked at each race's incarceration rates (per 100,000) across all 51 states.

Looking at the West region, from the data, New Mexico had the most Caucasian people incarcerated (902), Colorado had the most African American people incarcerated (3651), and Arizona had the most Native Hawaiian (8136) and Hispanic people incarcerated (1453).

When looking at the Northeast region, from the data, Pennsylvania had the most Caucasian people incarcerated (398), African American people incarcerated (3269), and Hispanic people incarcerated (2001), while Connecticut had the most Native Hawaiian people incarcerated (3431).

When looking at the Midwest region, from the data, Indiana had the most Caucasian people incarcerated (564), South Dakota had the most African American people incarcerated (4663) and Hispanic incarcerated (1727), and Wisconsin had the most Native Hawaiian people incarcerated (4817).

When looking at the South region, from the data, Oklahoma had the most Caucasian people incarcerated (889), West Virginia had the most African American people incarcerated (7360) and Native Hawaiian people incarcerated (3271), and Mississippi had the most Hispanic people incarcerated (4869).

With this being the case, we wonder why these states incarcerate the most out of their region. More specifically analyzation can go into the states Arizona, Pennsylvania, South Dakota, and West Virginia because they have a higher incarceration rate in more than one race, compared to other states in their region. Research can go into the laws for these states, the amount of people per race in these states, and possibly what crimes people are being incarcerated for.

On the other hand, looking at the West region, from the data, Alaska had the least Caucasian people incarcerated (383), Hawaii had the least African American people incarcerated (1032) and least Hispanic people incarcerated

(415), and Montana had the least Native Hawaiian people incarcerated (0).

Looking at the Northeast region, from the data, Vermont had the least Caucasian people incarcerated (226), Massachusetts had the least African American people incarcerated (1502), Rhode Island and Vermont had the least Native Hawaiian people incarcerated (0), and Maine had the least Hispanic people incarcerated (407).

Looking at the Midwest region, Minnesota had the least Caucasian people incarcerated (251), North Dakota had the least African American people incarcerated (1671), Nebraska had the least Native Hawaiian people incarcerated (156), and Illinois had the least Hispanic people incarcerated (472).

Looking at the South region, District of Columbia had the least Caucasian people incarcerated (88) and the least African American people incarcerated (1031), Alabama had the least Native Hawaiian people incarcerated (98), and Delaware had the least Hispanic people incarcerated (250).

Research can go into looking at states with the least amount of race(s) incarcerated and make correlations to states with the highest amount of race(s) incarcerated. Correlations such as education levels, immigration policies, and hardship with disenfranchised communities can help create new data in relations with incarceration rates across the country.

References

[1] A.W. Goodman. *On sets of acquaintances and strangers at any party. American Mathematics Monthly.* **66**, 778–783, (1959).

[2] K. Buschur. *Introduction to Ramsey Theory.* @inproceedingsBuschur2010INTRODUCTIONTR, (2010).

[3] M. Pawluk, and M. A. Waddell. *Using Ramsey Theory to Measure Unavoidable Spurious Correlations in Big Data. Axioms*, **8**(1), 29 (2019).