****

**The Relationship Between Students’ Physical Fitness and Rate of Crime in California:**

**A Multivariate Analysis**

**Prepared by**

*Aman Panwar,*

*Praveen Kumar, Jarrod Graham,*

**1. Problem Context**

Physical inactivity has been responsible for six percent of the global mortality and it has been ranked as the fourth leading risk factor of death (World Health Organization, 2010). For the age group of 5−17 years, physical activity provides essential health benefits for children and youth (Janssen, 2007; Janssen & Leblanc, 2010; Physical Activity Guidelines Advisory Committee, 2008). In addition, the behaviors related to health during the early teenage years between 11 and 16 years are crucial because the patterns of the health-related habits in their adulthoods emerge at this time (Gregory & Lowe, 2000).

In this context, the importance of students’ health and the necessity to figure out of the factors influencing the students’ physical fitness have been emphasizing. Safe neighborhoods that are free of crime and violence are an integral component of healthy neighborhoods. In addition to direct physical and mental impacts of crime, fear of crime and violence inhibit the use of community assets for physical exercise at nearby parks and playgrounds or walking or bicycling to commute to local destinations for basic needs (Fowler et al., 2009; Takagi, Ken’ichi, & Kawachi, 2012).

In this study, we decided to use multivariate statistical techniques to obtain deeper insights into the relationship between the rate of crime of an area and physical fitness of students living there.

Physical fitness and crime datasets were used for this study.

Physical fitness dataset has physical fitness score for students who attended grades five, seven, and nine in California public schools from 1998 to 2018.

Crime dataset has the rate of crime which represents the number of violent crimes per 1,000 population. Four types of crimes (murder, rape, robbery, and assault) and the total number of crimes are reported from years 2000 to 2013 at the geographical levels of city/towns, counties, regions, and state.

**2. Data Cleaning and Visualization**

From both data sets, it was appropriate to subset certain parts of the data to simplify our analysis in order to perform the novel statistical/visualization techniques that we didn’t apply in the previous study.

For the physical fitness data set, the cleaning methods that were applied included changing the names of columns to be consistent between data sets (specifically county\_name), subsetting certain variables from each of the data sets (removing unnecessary data for our study), aggregating the data by all students (instead of performing the study on a more granular level), filling missing values with column median rates (physical fitness), and then calculating rates of crime and physical fitness rates from this aggregated view of the data.

In original physical fitness dataset, the physical fitness was reported for students from Grades 5, 7, and 9 for years from 1998-2018 for every county. Similarly, crime rates were reported for all counties from year 2000-2013. Since the crime dataset had data for only 14 years, data for all years except 2000-2013 was removed from the physical fitness dataset. After this, physical fitness level of all students from all grades and years were averaged and an **average physical fitness level of counties, represented by avg\_percent** was reported.

Similarly, the **average crime rate represented by ‘avg\_rate’** for every county was reported in crime dataset. Crime dataset had all types of crime reported in one column so data in this column was separated to get different types of crimes in different columns. ‘Dcast’ and ‘summarize’ functions were utilized to perform these operations in R.

In the original crime dataset, the crime rate was reported by calculating the number of crimes per 1000 population. ‘*Violent crime\_total*’ in this dataset represents ‘*forcible rape*’, ‘*aggravated assault*’, ‘*murder*’, and ‘*robbery*’. The original crime and physical fitness dataset had 49,227 rows and 35,257 rows in them respectively and **the merged fitness-crime dataset of California state had 7 variables for different crime rates and physical fitness level for 58 counties in 58 rows.**

In order to produce more accurate results that are more representative of California as a whole, an outlier analysis was also performed. The method that was chosen to visualize what outliers needed to be removed was the bivariate boxplot. Fig. 1 below is an image of the bivariate boxplot that was produced.

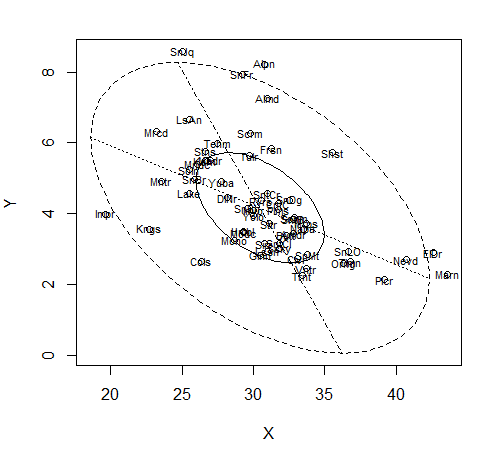


Figure 1: Bivariate boxplot.

The plot above indicates that there were 5 outlier counties that needed to be removed. In removing the outliers, results were expected to be more representative of the dataset and therefore producing more reliable results, which increases the reliability of drawn conclusions from this analysis. The correlation matrix of the merged fitness-crime data before and after removing the outliers is shown in Tabs. 1 and 2, respectively.

Table 1: Correlation matrix of the cleaned fitness-crime data, before removing outliers.

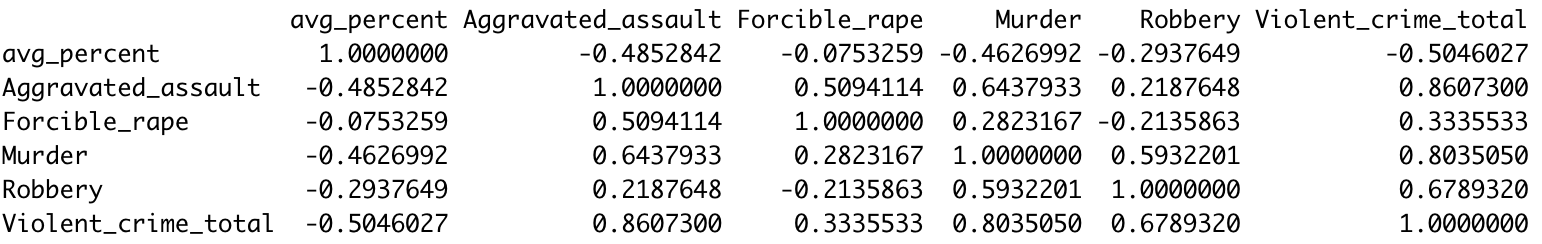
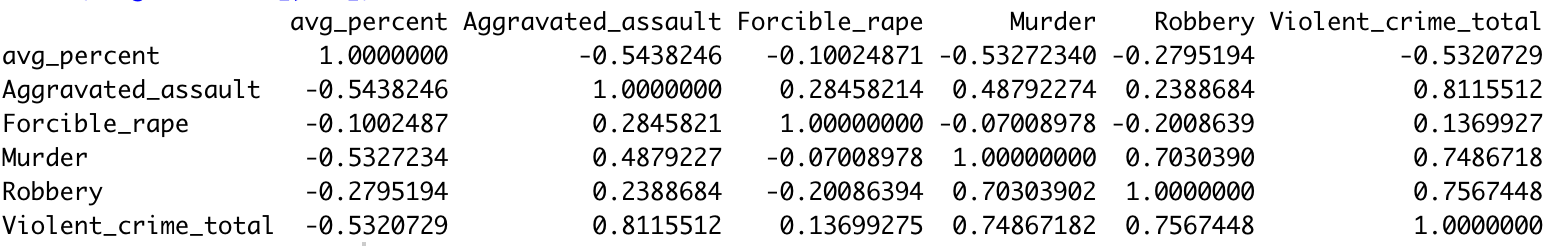


Table 2: Correlation matrix of the cleaned fitness-crime data, after removing outliers.



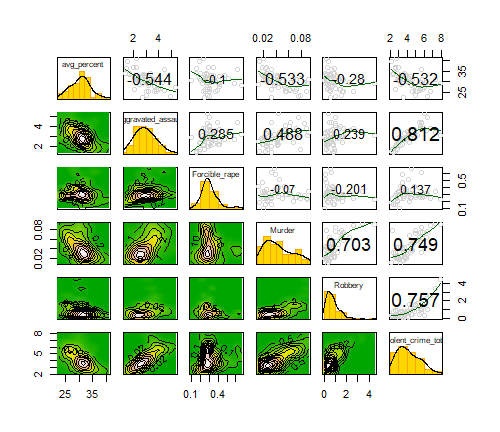


Figure 2: Kdepair plot of cleaned fitness-crime data.

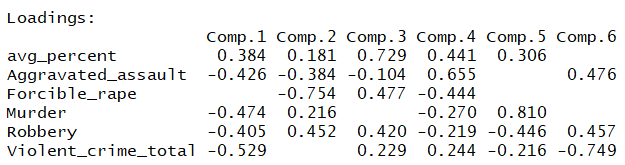
A kdepair plot (Fig. 2) for the fitness-crime dataset showed that there are mostly 2 clusters and sometimes just one cluster, for example, a 2D panel of *Forcible\_rape* and *Murder* clearly showed that there are two clusters in them. This plot not only helped us in exploring distribution in our data but also helped identify the number of clusters later in clustering section.

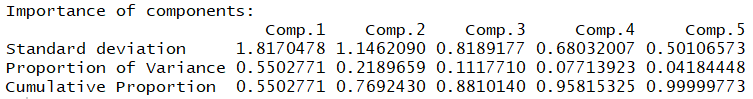
**3. Dimension Reduction**

There are 6 continuous variables in our dataset and it can be difficult if someone needs to derive insight from this dataset with several variables. To help overcome this problem, PCA (Principal Component Analysis) was performed to explain the dataset comprehensively with minimum possible variables.

PCA was performed on the fitness-crime dataset with outliers removed from it. The first two principal components (PC) represents 77% of the total variance and it is therefore enough to use first two PC to represent the original dataset. Based on their respective loadings, PC 1 can be thought to represent all crimes except *Forcible\_rape* and *Robbery* and PC 2 can be thought to represent *Forcible\_rape* and *Robbery*. A simple interpretation of PC1 score could be that county with higher PC1 score will have lower crime rates especially for the crimes with higher absolute magnitude of PC loadings and higher physical fitness level.

The results of PCA are as follows:





A PCA biplot (Fig. 3) between first two PC scores reveals that robbery and murder are positively correlated and similarly aggravated assault and forcible rape are positively correlated as the angles between correlated vectors is less than 90 degrees. Forcible rape and robbery are almost uncorrelated as they are orthogonal to each other. Physical fitness level represented by avg\_percent is negatively correlated with the crimes in general as the angle between them is greater than 90 degrees. This further supports our original hypothesis about the crime negatively affecting the fitness level of students.

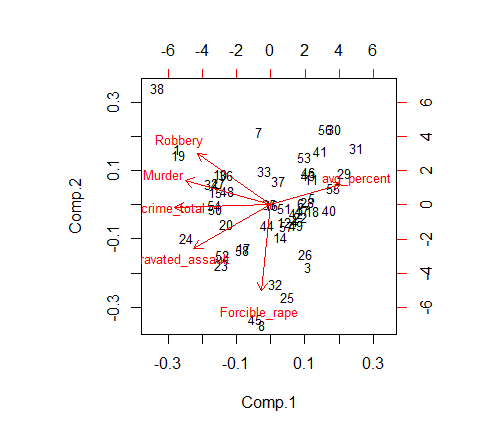


Figure 3: PCA biplot.

**4. Cluster Analysis**

We performed cluster analysis for discovering all possible clusters of most possible homogeneous observations in our fitness-crime dataset. We considered the following three clustering methods:

***Hierarchical Clustering (HC):***

In HC, data is not partitioned into particular groups in a single step. The process requires a series of partitions that can run from a single cluster containing all individuals to *n* clusters, each containing a single individual. This clustering method uses the distance matrix (standardized Euclidean distance) and is based on the choices of distance between groups further divided into three categories:

*Single linkage Hierarchical clustering:*

Single linkage HC uses the smallest distance between the two groups. The single linkage dendrogram of cleaned fitness-crime dataset is shown in Fig. 4 and corresponding scree plot of the inter-cluster distances is shown in Fig. 5, which shows two major probable clusters in fitness-crime data.

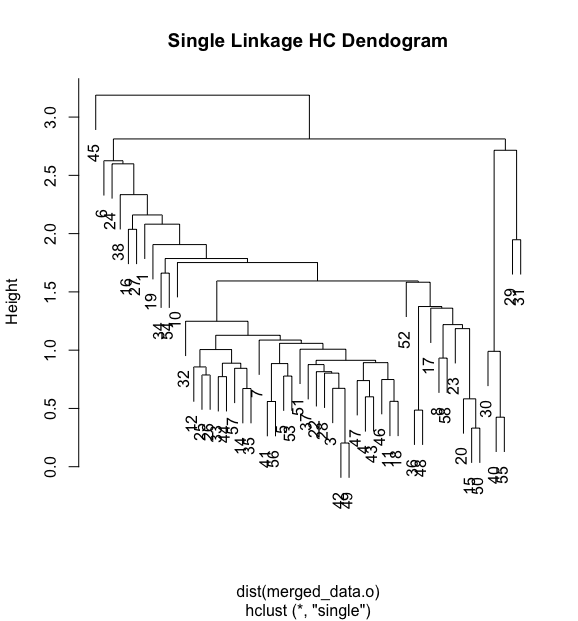
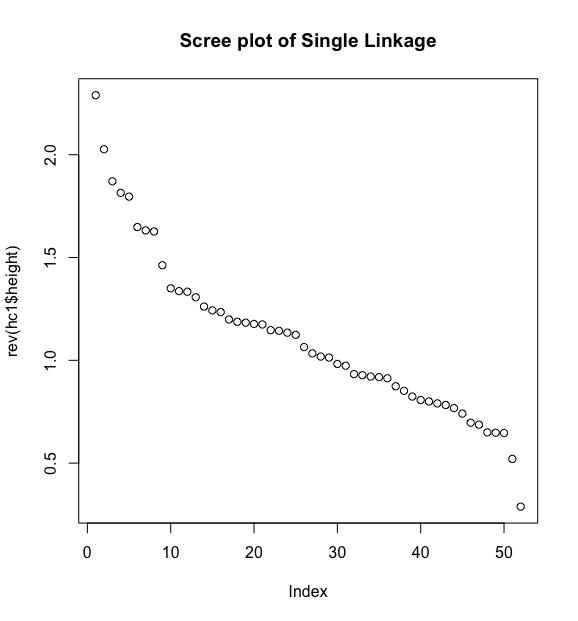


Figure 4: Hierarchical clustering single linkage dendrogram.

**Figure 5: Hierarchical clustering single linkage scree plot.

*Complete linkage Hierarchical clustering:*

Complete linkage HC uses a maximum distance between the two groups. The complete linkage dendrogram of cleaned fitness-crime dataset is shown in Fig. 6 and corresponding scree plot of the inter-cluster distances is shown in Fig. 7, which shows four probable clusters in fitness-crime data.

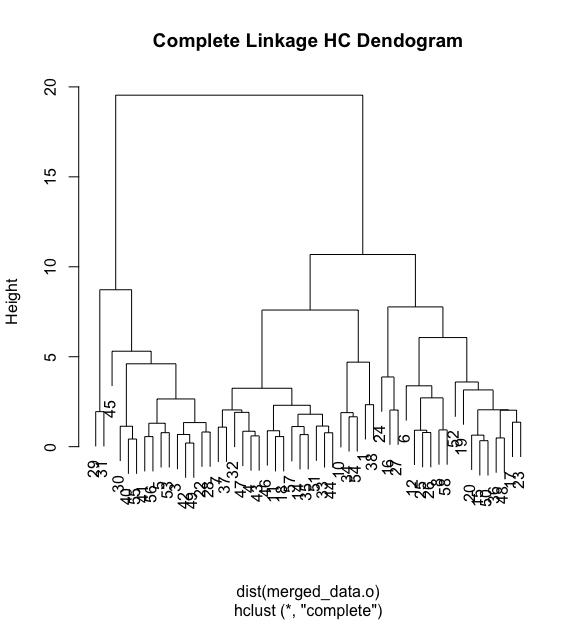


Figure 6: Hierarchical clustering complete linkage dendrogram.

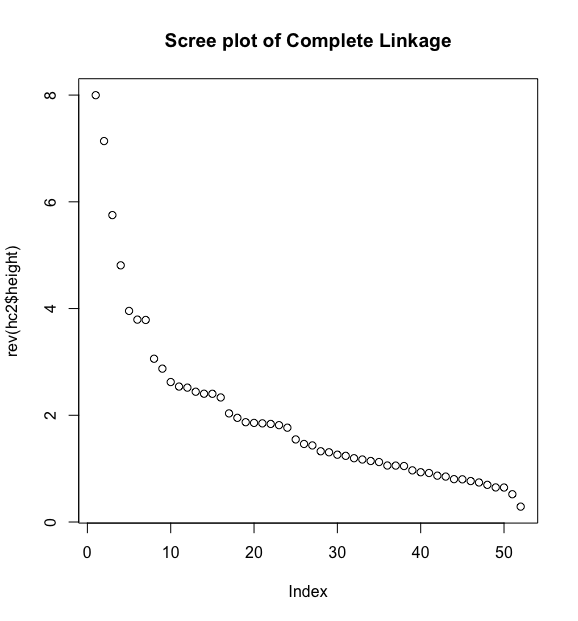


Figure 7: Hierarchical clustering complete linkage scree plot.

*Average linkage Hierarchical clustering:*

Average linkage HC uses average distance between all possible pairs. The average linkage dendrogram of cleaned fitness-crime dataset is shown in Fig. 8 and corresponding scree plot is shown in Fig. 9, which shows two clusters in fitness-crime data. HC clustering plot is shown in Fig. 10.

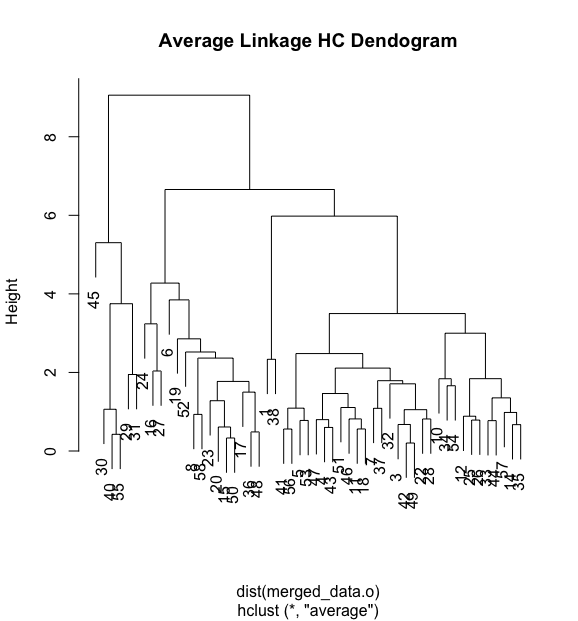


Figure 8: Hierarchical clustering average linkage dendrogram.

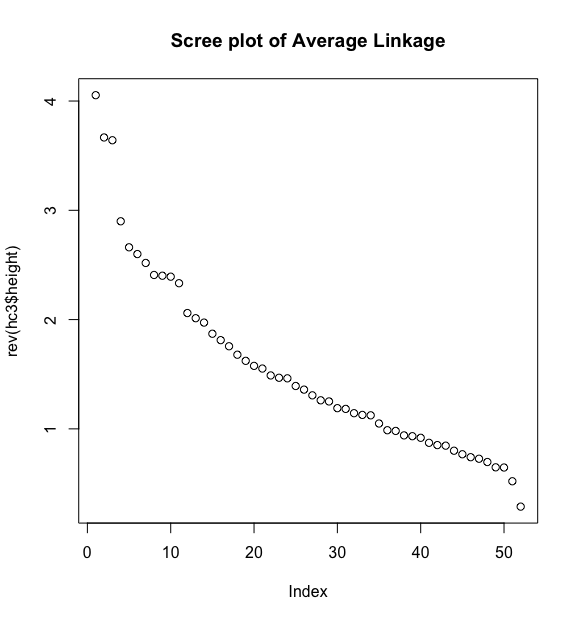


Figure 9: Hierarchical clustering average linkage scree plot.

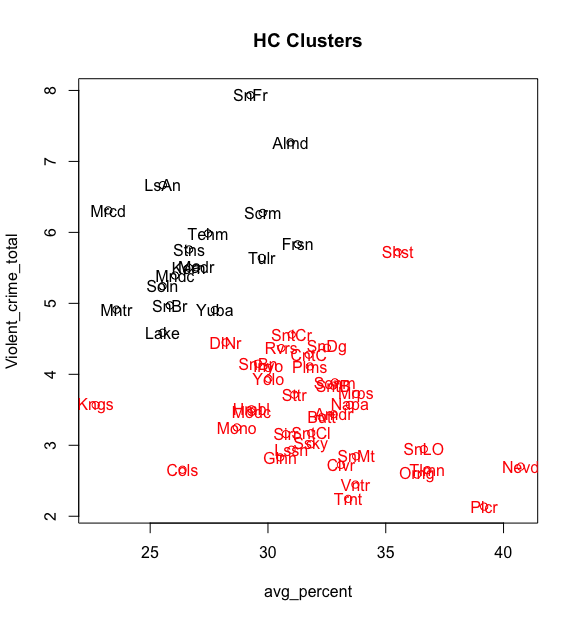
******

Figure 10: Hierarchical clusters.

***4.2 K-Means (KM) Clustering:***

Next we consider K-means clustering which tries to find a partition of the *n* observations into *k* groups that minimizes the within-group sum of squares (WGSS) over all variables. In KM, the input is data and uses a random initial assignment that sometimes provide different clustering result even for the same data set. KM clustering scree plot of fitness-crime data is shown in Fig. 11, which shows two major clusters in fitness-crime data.

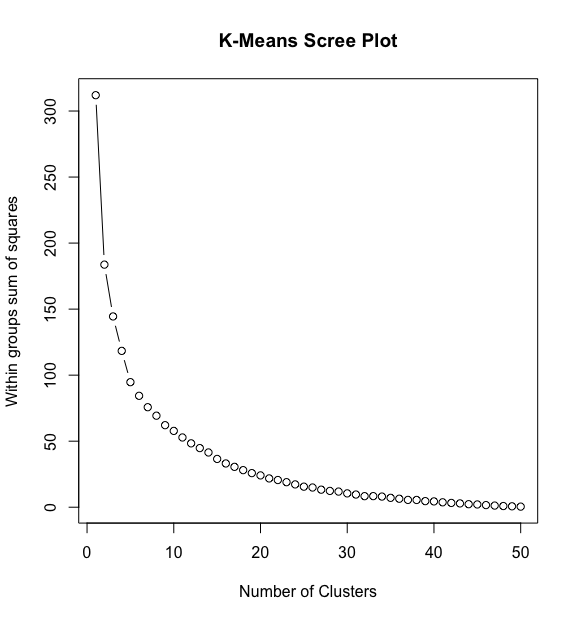


Figure 11: K-Means clustering scree plot.

***4.3 Model-based Clustering:***

Finally, we applied model-based clustering to perform cluster analysis of the data. This clustering method can model different patterns and hence offers a very flexible and powerful clustering procedures which is based on finite mixture model. The BIC (Bayes Inf. Criteria) plot of cleaned fitness-crime data is shown in Fig. 12. The maximum BIC is used as a measure of the number of clusters, which shows **two** clusters in fitness-crime data. First cluster has 45 counties and in second there are 8 counties of the California state.

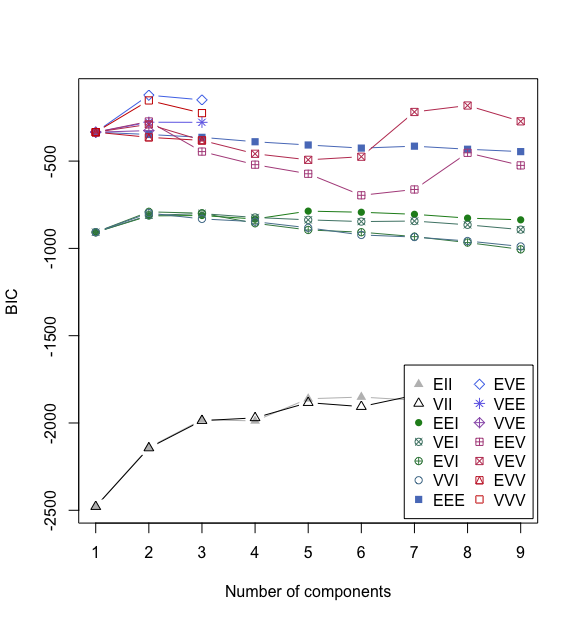
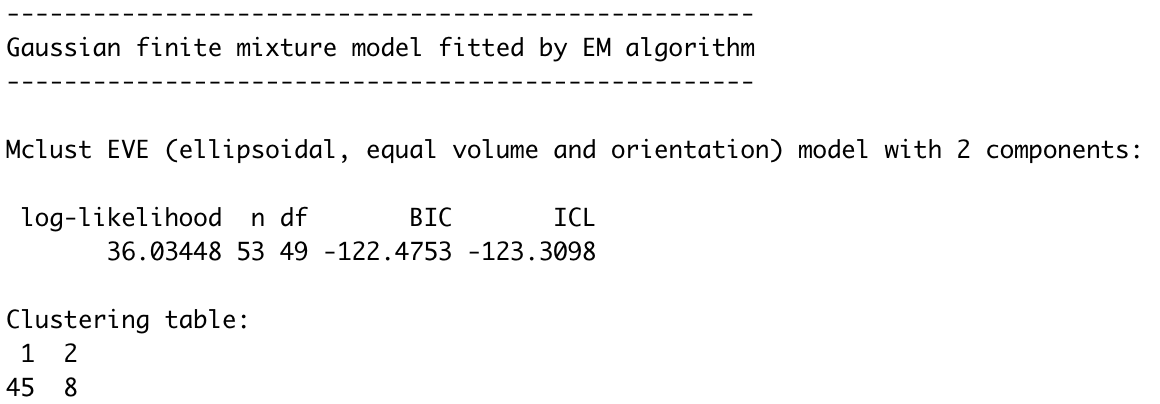


Figure 12: Model-based clustering BIC plot.

Results of model-based clustering are as follows:



We eventually concluded that **hierarchical based clustering method is the best method to segregate our observations into two clusters** because it was evident in its plot that there was clear separation between the observations from two clusters. Although, in the absence of information on true clusters, it is difficult to be completely confident if the data has been well segregated.

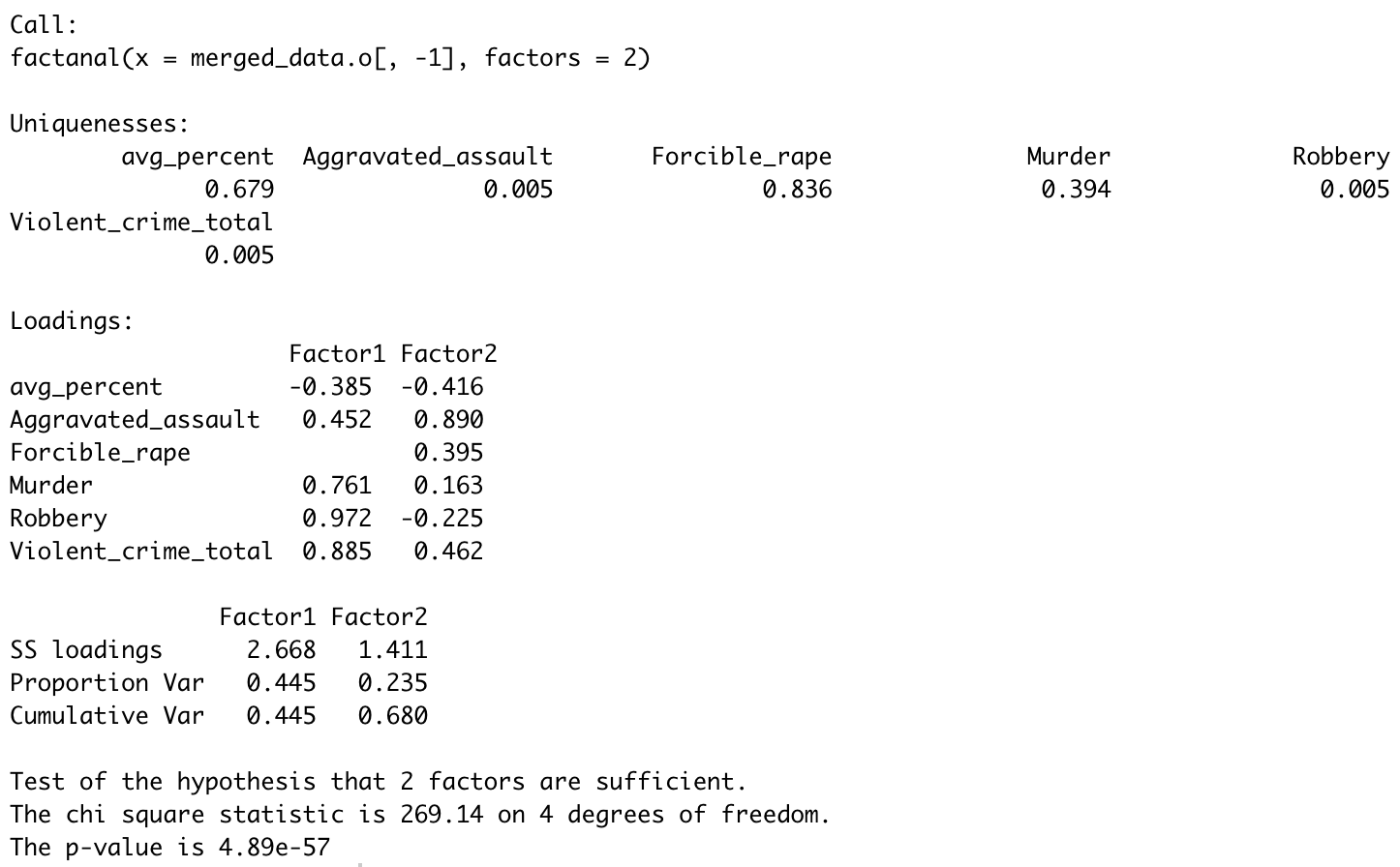
**5. Factor Analysis**

The basis of factor analysis is a regression model linking the manifest variables (data) to a set of unobservable latent variables (common factors). For our fitness-crime data,we performed both the exploratory factor analysis (EFA) and confirmatory factor analysis (CFA).

***5.1 Exploratory Factor Analysis:***

EFA investigates relationship between manifest and latent variables without making any assumptions about which manifest variable is related to which factor.

Results of EFA analysis on fitness-crime data are as follows:

****

EFA results show very high values of uniqueness for *avg\_percent* (=0.68), *Forcible\_rape* (=0.84), *Murder* (=0.39) and very small values (=0.005) for all of the remaining variables, *Aggravated\_assult*, *Robbery*, and *Violent\_crime\_total*, respectively. A two factor model is sufficient for our fitness-crime data, where *Murder*, *Robbery*, and *Violent\_crime\_total* contributing strongly (loadings > 0.76) to “Factor1”, and *avg\_percent*, *Aggravated\_assult*, and *Forcible\_rape* (loadings >= 0.40) to “Factor2”, respectively. Though, there is not a big difference in *avg\_percent* loadings for Factor1 and Factor2.

We compared statistical model correlation matrix with the correlation matrix of the original fitness-crime data by measuring the root-mean-square error (RMSE). The RMSE value obtained is **0.045**, which is less than 0.05. This confirms the appropriate number of factors is 2, and hence fail to reject the null hypothesis.

Based on the information about variables and loading coefficients, we can interpret factors as follows:

* Factor 1 can be of type say Crime1.
* Factor 2 can be of type Crime2.

***Confirmatory Factor Analysis:***

CFA is used to test whether a specific factor model provides an adequate fit for the covariances/correlations between the manifest variables. First, we build a model(s) (not listed in the report) suggested by EFA analysis to perform CFA. All of such models failed to compute coefficient of covariances.

Finally, we proposed the following model based on two factors---**Fitness** and **Crime**--- as given below:

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*Model(text = "*

*Fitness -> avg\_percent, NA, 1*

*Crime -> Aggravated\_assault, lambda2, NA*

*Crime -> Forcible\_rape, lambda3, NA*

*Crime -> Murder, lambda4, NA*

*Crime -> Robbery, lambda5, NA*

*Crime -> Violent\_crime\_total, lambda6, NA*

*Fitness <-> Crime, rho, NA*

*avg\_percent <-> avg\_percent, theta1, NA*

*Aggravated\_assault <-> Aggravated\_assault, theta2, NA*

*Forcible\_rape <-> Forcible\_rape, theta3, NA*

*Murder <-> Murder, theta4, NA*

*Robbery <-> Robbery, theta5, NA*

*Violent\_crime\_total <-> Violent\_crime\_total, theta3, NA*

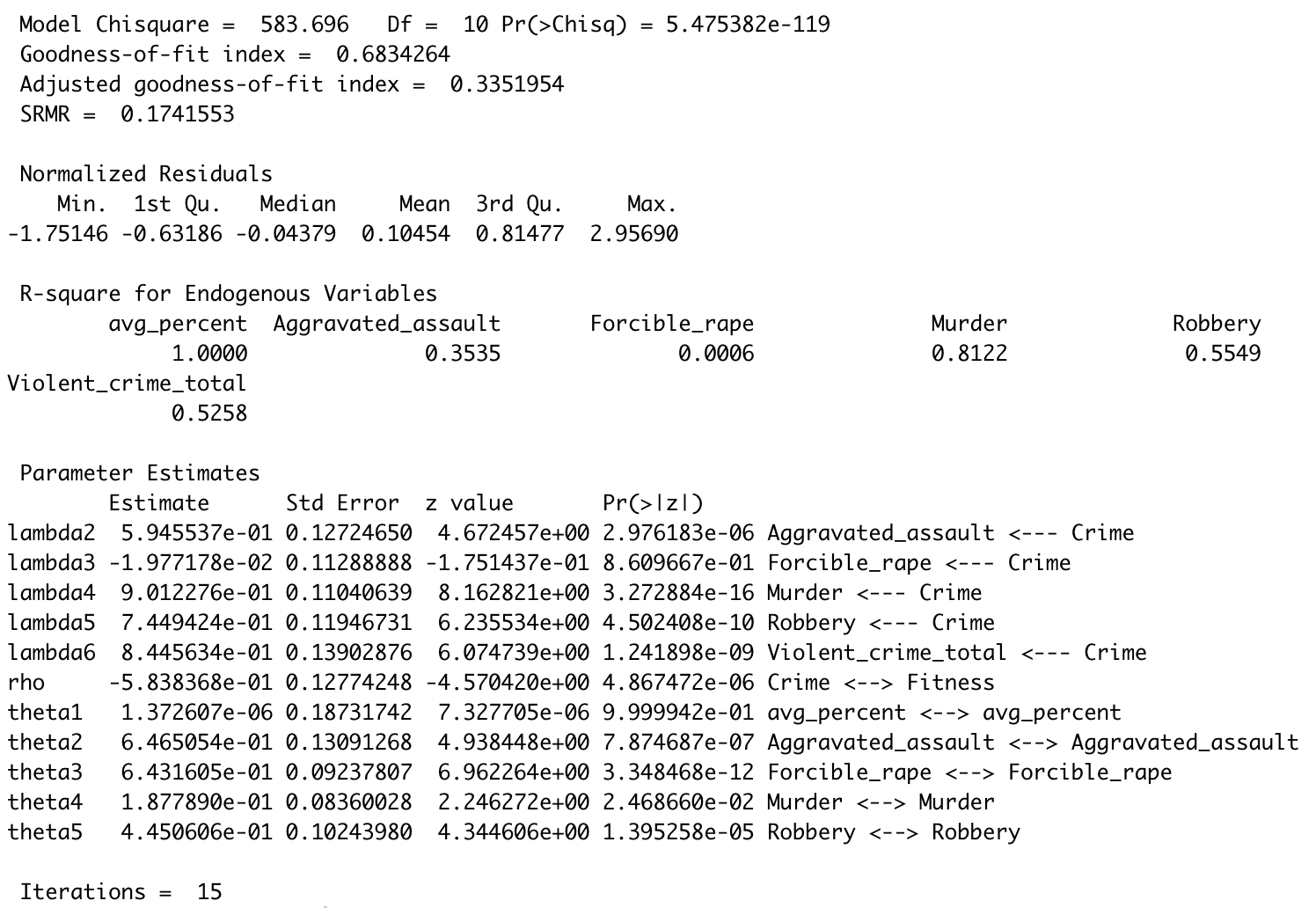
*Fitness <-> Fitness, NA, 1*

*Crime <-> Crime, NA, 1*

*")\_*

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

Results of CFA analysis are as follows:



The values obtained for the Goodness-of-fit index (GFI = 0.68) and Adjusted Goodness-of-fit index (AGFI = 0.34) are less than 0.95, and the standard root mean square difference (SRMR = 0.17) value is greater than 0.05, which implies that the fitness-crime data does not support the CFA model.

The path diagram of the model is shown in Fig. 13, which shows strong negative correlation (= -0.54) between the **Fitness** and **Crime** factors. In our model, we choose fitness loading value equal to 1.

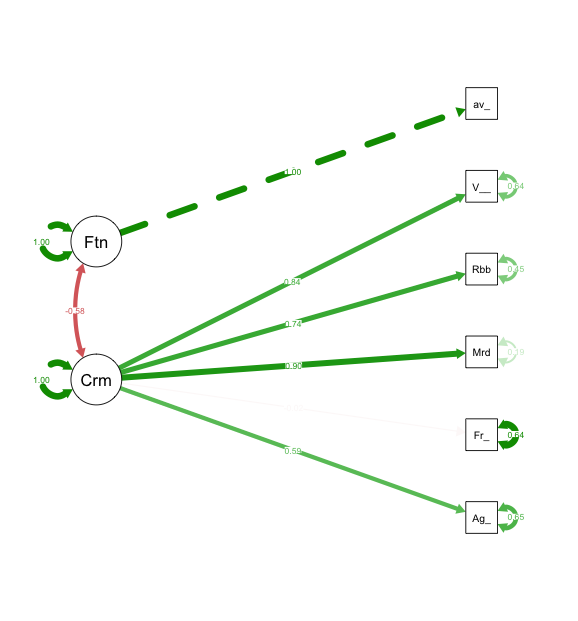


Figure 13: Path diagram of the CFA model.

**6. Conclusions and Future Directions**

In continuing our investigation of the impact of rate of crime on the physical fitness of students in California after our first study, we were able to achieve a deeper understanding of this problem. Insights that were obtained from the analyses were visualizing 2 clusters, labeling 2 principal components, determining correlations between variables already present in the data set, and identifying 2 latent variables within the data.

The PCA analysis yielded 2 principal components that were able to explain about 77% of the original data. It was reasonable to label the PC 2 as representing *Forcible\_rape* and *Robbery* while PC 1 likely represented all other crimes present in the analysis. With this, we were able to simplify the view of the data and offer a new perspective of what types of crimes can be considered related in some capacity. Additionally, the PCA also yielded correlations between variables that were positive between *Robbery* and *Murder*, and *Forcible\_rape* and *aggravated\_assault*, which are sensible results because the crimes can be logically related to each other.

Cluster analysis obtained by using HC, K-means, and model-based techniques confirm two main clusters of California counties in fitness-crime data, one for the lower crime rate counties and second for the higher crime rate counties. Using CFA we observed two common factors ---**Fitness** and **Crime**---in fitness-crime data, which are partially derived from EFA. However, CFA does not confirm this model and showed a negative correlation between these factors.

The identified latent variables were labeled Crime 1 and Crime 2. The reason for this is because we could see that the associated crimes (Crime1 = *Robbery* and *Murder*, Crime2 = *aggravated\_assault* and *Forcible\_rape*) are difficult to logically relate to each other. One way we could attempt to reach for a label other than Crime1 and Crime2 could be to propose that Factor1 includes crimes likely committed with a weapon (*Robbery* and *Murder*). Factor2 could be thought of crimes that are committed without a weapon.

After the analyses performed in this study, we determined some additional dimensions that would be interesting to investigate. Two dimensions that were already available in the original datasets were race and gender. Essentially, it is possible to investigate the impact of rate of crime on the physical fitness between males and females and determine if there is a tendency for the amount of crime to impact one more than the other. Regarding race, it is also possible to determine if higher or lower rates of crime impact the physical fitness of different races more than others. A dimension that was not available in the original datasets could include exploring the impact of different rates of crimes on people of different income groups. That is, the point of focus would be if varying rates of crime impact people of varying income groups differently.

**References:**

● Fowler, P. J. et al. (2009). Community violence: A meta-analysis on the effect of exposure and mental health outcomes of children and adolescents. *Dev Psychopathol, 21*(1), 227−259.

● Takagi, D., Ken'ichi, I., & Kawachi, I. (2012). Neighborhood social capital and crime victimization: Comparison of spatial regression analysis. *Soc Sci Med.* *75*(10), 1895−902.

**Supporting Information:**

# Percentage of Physically-fit Students (LGHC Indicator)

#<https://healthdata.gov/dataset/percentage-physically-fit-students-lghc-indicator>

# Physical Fitness Testing (PFT)

# <https://www.cde.ca.gov/ta/tg/pf/>

# Student Poverty FRPM Data

# <https://www.cde.ca.gov/ds/sd/sd/filessp.asp>

# Public Schools and Districts Data Files

# <https://www.cde.ca.gov/ds/si/ds/pubschls.asp>

# Violent Crime Rate

# <https://data.ca.gov/dataset/violent-crime-rate>