Proyecto de Estadística Fase 2

Enrique Martínez González ENRIQUE.MARTINEZ@ESTUDIANTES.MATCOM.UH.CU
Grupo C412

Osmany Pérez Rodríguez OSMANY.PEREZ@ESTUDIANTES.MATCOM.UH.CU
Grupo C412

Carmen Irene Cabrera Rodríguez CARMEN.CABRERA@ESTUDIANTES.MATCOM.UH.CU
Grupo C412

1. Introducción

El set de datos a utilizar contiene información sobre comunidades de Estados Unidos, que puede estar relacionada de alguna forma a los índices de criminalidad que estas presentan. Se combinan datos socioeconómicos del censo de 1990, ley datos de aplicación de la encuesta de 1990 sobre gestión de la aplicación de la ley y estadísticas administrativas, y datos delictivos de la UCR del FBI de 1995.

El propósito de este proyecto es llevar a cabo un estudio sobre las variables de dicho juego de datos, aplicando técnicas de regresión, reducción de dimensión y anova que permitan luego realizar una caracterización de las mismas, así como la realización de predicciones sobre ellas.

2. Variables

El set de datos que se analiza en este proyecto tiene como uno de los objetivos principales el análisis de la criminalidad en todas las "municipalidades" en los Estados Unidos. Se decidió hacer un análisis tratando de asociar cuestiones meramente socio-ecónomicas de los individuos en búsqueda de una relación de los mismos con el crimen en estas comunidades.

Como consecuencia a dicha decisión se toman como variables independientes a analizar:

- medIncome : ingreso promedio de los trabajadores.
- PctPopUnderPov : porciento de la población que vive por debajo del nivel de pobreza.
- PctUnemployed : porciento de individuos desempleados con edad laboral.

Como variables dependientes se toman:

- ViolentCrimesPerPop : cantidad de crímenes violentos por cada 100 000 habitantes.
- nonViolPerPop : cantidad de crímenes no violentos por cada 100 000 habitantes.

3. Regresión lineal

```
> cor(dataFiltered)
                    medIncome PctPopUnderPov PctUnemployed ViolentCrimesPerPop nonViolPerPop
medIncome           1.0000000    -0.7583720    -0.6146940          -0.3761265    -0.4660912
PctPopUnderPov     -0.7583720     1.0000000     0.7727260           0.4743996     0.5061137
PctUnemployed      -0.6146940     0.7727260     1.0000000           0.4317157     0.3985916
ViolentCrimesPerPop -0.3761265    0.4743996     0.4317157           1.0000000     0.6289104
nonViolPerPop      -0.4660912     0.5061137     0.3985916           0.6289104     1.0000000
```

Existe una dependencia lineal entre medIncome y PctPopUnderPov. Luego es posible eliminar una de las variables, pero por el momento se continuará trabajando con todas en la construcción del modelo.

Se tomará como variables dependientes a ViolentCrimesPerPop y nonViolPerPop.

Crímenes Violentos.

- ViolentCrimesPerPop→medIncome+ PctPopUnderPov+PctUnemployed

```
> multi.fit = lm(ViolentCrimesPerPop~medIncome+PctPopUnderPov+PctUnemployed)
> summary(multi.fit)

Call:
lm(formula = ViolentCrimesPerPop ~ medIncome + PctPopUnderPov +
    PctUnemployed)

Residuals:
    Min      1Q  Median      3Q     Max
-1804.7  -268.9  -105.8   172.3  3764.7

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    175.465468  65.690427   2.671  0.00762 **
medIncome       -0.001204   0.001242  -0.969  0.33244
PctPopUnderPov  22.380069   2.409455   9.288  < 2e-16 ***
PctUnemployed   32.168877   5.913909   5.440 5.93e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 510.2 on 2211 degrees of freedom
Multiple R-squared:  0.2359,    Adjusted R-squared:  0.2349
F-statistic: 227.5 on 3 and 2211 DF,  p-value: < 2.2e-16
```

El estimado del coeficiente del intercepto es 0 y no posee una diferencia circunstancial con los demás coeficientes. Tiene un nivel de significación muy bajo ya que $Pr(>|t|) = 1$. Se puede decir que las variables independientes definen a *ViolentCrimesPerPop*. El nivel de significación es 0 para las variables, excepto para medIcome que es 0.332. Esto apunta a que dicha variable no aporta nada al modelo y por tanto puede ser eliminada. El valor de R^2 ajustado es 0.2349 lo cual es una clara indicación de que el modelo es muy malo. El p-valor de F es 0, lo que significa que hay al menos una variable con valor significativamente mayor que cero.

- ViolentCrimesPerPop→PctPopUnderPov +PctUnemployed

©2021 E. Martínez, O. Pérez, C.Cabrera.

```
> multi.fit = lm(ViolentCrimesPerPop~PctPopUnderPov+PctUnemployed)
> summary(multi.fit)

Call:
lm(formula = ViolentCrimesPerPop ~ PctPopUnderPov + PctUnemployed)

Residuals:
    Min      1Q  Median      3Q     Max
-1844.9  -269.9  -107.3   174.5  3760.1

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   116.776     25.496    4.58 4.90e-06 ***
PctPopUnderPov 23.702      1.986   11.93  < 2e-16 ***
PctUnemployed  32.566      5.900    5.52 3.79e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 510.2 on 2212 degrees of freedom
Multiple R-squared:  0.2356,    Adjusted R-squared:  0.2349
F-statistic: 340.9 on 2 and 2212 DF,  p-value: < 2.2e-16
```

El estimado del coeficiente del intercepto es 0 y no posee una diferencia circunstancial con los demás coeficientes. Tiene un nivel de significación muy bajo ya que $Pr(>|t|) = 1$. Se puede decir que las variables independientes definen a ViolentCrimesPerPop. El nivel de significación es 0 para las variables, lo cual representa que tienen gran importancia. El valor de R^2 ajustado es 0.2349 lo cual es una clara indicación de que el modelo es muy malo. El p-valor de F es 0, lo que significa que hay al menos una variable con valor significativamente mayor que cero.

- ViolentCrimesPerPop→PctUnemployed

```
> multi.fit = lm(ViolentCrimesPerPop~PctUnemployed)
> summary(multi.fit)

Call:
lm(formula = ViolentCrimesPerPop ~ PctUnemployed)

Residuals:
    Min      1Q  Median      3Q     Max
-2190.2  -305.2  -129.2   168.7  3735.3

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    63.355     25.889   2.447   0.0145 *
PctUnemployed  86.965      3.862  22.515   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 526.3 on 2213 degrees of freedom
Multiple R-squared:  0.1864,    Adjusted R-squared:  0.186
F-statistic: 506.9 on 1 and 2213 DF,  p-value: < 2.2e-16
```

El estimado del coeficiente del intercepto es 0 y no posee una diferencia circunstancial con los demás coeficientes. Tiene un nivel de significación muy bajo ya que $Pr(>|t|) = 1$. Se puede decir que la variable independiente define a ViolentCrimesPerPop. El nivel de significación es 0 para la variable, lo cual representa que tiene gran importancia. El valor de R^2 ajustado es 0.186 lo cual es una clara indicación de que el modelo es muy malo. El p-valor de F es 0, lo que significa que la variable tiene valor significativamente mayor que cero.

- ViolentCrimesPerPop→PctPopUnderPov

```
> multi.fit = lm(ViolentCrimesPerPop~PctPopUnderPov)
> summary(multi.fit)

Call:
lm(formula = ViolentCrimesPerPop ~ PctPopUnderPov)

Residuals:
    Min      1Q  Median      3Q     Max
-1505.4  -267.0  -106.6   181.1  3852.4

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   215.191     18.347   11.73   <2e-16 ***
PctPopUnderPov  32.175      1.269   25.35   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 513.6 on 2213 degrees of freedom
Multiple R-squared:  0.2251,    Adjusted R-squared:  0.2247
F-statistic: 642.7 on 1 and 2213 DF,  p-value: < 2.2e-16
```

El estimado del coeficiente del intercepto es 0 y no posee una diferencia circunstancial con los demás coeficientes. Tiene un nivel de significación muy bajo ya que $Pr(>|t|) = 1$. Se puede decir que la variable independiente define a ViolentCrimesPerPop. El nivel de significación es 0 para la variable, lo cual representa que tiene gran importancia. El valor de R^2 ajustado es 0.2247 lo cual es una clara indicación de que el modelo es muy malo pero tiene mayor importancia que la variable independiente analizada previamente. El p-valor de F es 0, lo que significa que la variable tiene valor significativamente mayor que cero.

Crímenes no violentos.

- nonViolPerPop→medIncome +PctPopUnderPov+PctUnemployed

```
> multi.fit = lm(nonViolPerPop~medIncome+PctPopUnderPov+PctUnemployed)
> summary(multi.fit)

Call:
lm(formula = nonViolPerPop ~ medIncome + PctPopUnderPov + PctUnemployed)

Residuals:
    Min      1Q  Median      3Q     Max
-7287.5 -1349.4  -319.5  1028.1 23461.4

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.904e+03  2.945e+02  16.654  < 2e-16 ***
medIncome     -3.858e-02  5.568e-03  -6.928  5.6e-12 ***
PctPopUnderPov 1.108e+02  1.080e+01  10.260  < 2e-16 ***
PctUnemployed  4.502e+00  2.651e+01   0.170    0.865
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2287 on 2211 degrees of freedom
Multiple R-squared:  0.2721,    Adjusted R-squared:  0.2711
F-statistic: 275.5 on 3 and 2211 DF,  p-value: < 2.2e-16
```

El estimado del coeficiente del intercepto 0 y no posee una diferencia circunstancial con los demás coeficientes. Tiene un nivel de significación muy bajo ya que $Pr(>|t|) = 1$. Se puede decir que las variables independientes definen a nonViolPerPop. El nivel de significación es 0 para las variables, excepto para PctUnemployed que es 0.865. Esto apunta a que dicha variable no aporta nada al modelo y por tanto puede ser eliminada. El valor de R^2 ajustado es 0.2711 lo cual es una clara indicación de que el modelo es muy malo. El p-valor de F es 0, lo que significa que hay al menos una variable con valor significativamente mayor que cero.

- nonViolPerPop→medIncome+PctPopUnderPov

```
> multi.fit = lm(nonViolPerPop~medIncome+PctPopUnderPov)
> summary(multi.fit)

Call:
lm(formula = nonViolPerPop ~ medIncome + PctPopUnderPov)

Residuals:
    Min      1Q  Median      3Q     Max
-7278.3 -1352.9  -317.7  1029.5 23464.0

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.921e+03  2.774e+02  17.737  < 2e-16 ***
medIncome   -3.864e-02  5.554e-03  -6.958 4.54e-12 ***
PctPopUnderPov  1.119e+02  8.669e+00  12.909  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2287 on 2212 degrees of freedom
Multiple R-squared:  0.2721,    Adjusted R-squared:  0.2714
F-statistic: 413.4 on 2 and 2212 DF,  p-value: < 2.2e-16
```

El estimado del coeficiente del intercepto 0 y no posee una diferencia circunstancial con los demás coeficientes. Tiene un nivel de significación muy bajo ya que $Pr(>|t|)$ =1. Se puede decir que las variables independientes definen a nonViolPerPop. El nivel de significación es 0 para las variables, lo cual representa que tienen gran importancia. El valor de R^2 ajustado es 0.2714 lo cual es una clara indicación de que el modelo es muy malo. El p-valor de F es 0, lo que significa que hay al menos una variable con valor significativamente mayor que cero.

- nonViolPerPop→medIncome

```
> multi.fit = lm(nonViolPerPop~medIncome)
> summary(multi.fit)

Call:
lm(formula = nonViolPerPop ~ medIncome)

Residuals:
    Min      1Q  Median      3Q     Max
-6010.6 -1518.2  -392.9  1076.0 22898.6

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.069e+03  1.371e+02   58.84   <2e-16 ***
medIncome   -9.301e-02  3.753e-03  -24.78   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2371 on 2213 degrees of freedom
Multiple R-squared:  0.2172,    Adjusted R-squared:  0.2169
F-statistic: 614.2 on 1 and 2213 DF,  p-value: < 2.2e-16
```

El estimado del coeficiente del intercepto 0 y no posee una diferencia circunstancial con los demás coeficientes. Tiene un nivel de significación muy bajo ya que $Pr(>|t|)$ =1. Se puede decir que la variable independiente define a ViolentCrimesPerPop. El nivel de significación es 0 para la variable, lo cual representa que tiene gran importancia. El valor de R^2 ajustado es 0.2169 lo cual es una clara indicación de que el modelo es muy malo. El p-valor de F es 0, lo que significa que la variable tiene valor significativamente mayor que cero.

- nonViolPerPop→PctUnemployed

```
> multi.fit = lm(nonViolPerPop~PctUnemployed)
> summary(multi.fit)

Call:
lm(formula = nonViolPerPop ~ PctUnemployed)

Residuals:
    Min      1Q  Median      3Q     Max
-6495.5 -1575.8  -395.8  1078.9 22814.6

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2678.90     120.90   22.16   <2e-16 ***
PctUnemployed  368.78      18.04   20.45   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2458 on 2213 degrees of freedom
Multiple R-squared:  0.1589,    Adjusted R-squared:  0.1585
F-statistic:   418 on 1 and 2213 DF,  p-value: < 2.2e-16
```

El estimado del coeficiente del intercepto 0 y no posee una diferencia circunstancial con los demás coeficientes. Tiene un nivel de significación muy bajo ya que $Pr(>|t|)$ =1. Se puede decir que la variable independiente define a ViolentCrimesPerPop. El nivel de significación es 0 para la variable, lo cual representa que tiene gran importancia. El valor de R^2 ajustado es 0.1585 lo cual es una clara indicación de que el modelo es muy malo además de tener menor importancia que la variable independiente analizada previamente. El p-valor de F es 0, lo que significa que la variable tiene valor significativamente mayor que cero.

Supuestos.

A pesar que ninguno de los modelos resultó ser siquiera medianamente buenos, se escogió un modelo de cada una de las secciones analizadas previamente. A continuación se realizará el análisis de los supuestos en cada caso. No se requiere hacerlos todos una vez que uno se incumple (contrario a lo que se presenta aquí).

- ViolentCrimesPerPop→PctPopUnderPov+PctUnemployed

```
> mean(multi.fit$residuals)
[1] 1.159935e-13
> sum(multi.fit$residuals)
[1] 2.606733e-10
```

La media de los errores es cero y la suma de los errores es cero.

```
> ks.test(res, 'pnorm', mean = mean(res),sd=sd(res))

        One-sample Kolmogorov-Smirnov test

data:  res
D = 0.12433, p-value < 2.2e-16
alternative hypothesis: two-sided
```

El p-valor es menor que 0.05 por tanto no se cumple que los errores siguen una distribución normal.

```
> dwtest(multi.fit)

        Durbin-Watson test

data:  multi.fit
DW = 1.6345, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

Como el p-valor de esta prueba es 0 ¡ 0.05 podemos rechazar la hipótesis nula por lo que podemos afirmar que los errores no son independientes.

```
> bptest(multi.fit)

        studentized Breusch-Pagan test

data:  multi.fit
BP = 267.61, df = 2, p-value < 2.2e-16
```

Como el p-valor de esta prueba es 0 ¡ 0.05 podemos rechazar la hipótesis nula por lo que podemos afirmar que no se cumple la heterocedasticidad. Por lo que el supuesto de Homocedasticidad no se mantiene.

- nonViolPerPop→medIncome+PctPopUnderPov

```
> mean(multi.fit$residuals)
[1] -1.122013e-14
> sum(multi.fit$residuals)
[1] -1.54472e-11
```

La media de los errores es cero y la suma de los errores es cero.

```
> ks.test(res, 'pnorm', mean = mean(res),sd=sd(res))

        One-sample Kolmogorov-Smirnov test

data:  res
D = 0.092453, p-value < 2.2e-16
alternative hypothesis: two-sided
```

El p-valor es menor que 0.05 por tanto no se cumple que los errores siguen una distribución normal.

```
> dwtest(multi.fit)

        Durbin-Watson test

data:  multi.fit
DW = 1.6697, p-value = 3.454e-15
alternative hypothesis: true autocorrelation is greater than 0
```

Como el p-valor de esta prueba es 0 ¡ 0.05 podemos rechazar la hipótesis nula por lo que podemos afirmar que los errores no son independientes.

```
> bptest(multi.fit)

        studentized Breusch-Pagan test

data:  multi.fit
BP = 33.516, df = 2, p-value = 5.273e-08
```

Como el p-valor de esta prueba es 0 ¡ 0.05 podemos rechazar la hipótesis nula por lo que podemos afirmar que no se cumple la heterocedasticidad. Por lo que el supuesto de Homocedasticidad no se mantiene.

Además de ser modelos, que de manera previa a esta comprobación, se conocía que eran malos; incumplen varios de los supuestos lo que reafirma que no son válidos.

4. Reducción de dimensión

Nuestro set de datos posee un total de 145 variables, el cual es un número bastante grande para intentar graficar todas las variables para apreciar a simple vista si los datos están correlacionados o no, de igual forma la matriz de correlación sigue siendo bastante grande pero podemos ver que muy pocos datos tienen correlación entre si. Intentando disminuir la cantidad de variables realizamos ACP pero dado los resultados anteriores esta no es altamente correlacionada por lo que habría que tomar un número igual grande de componentes principales para poder representar un % no tan grande de la muestra.

Intentemos agrupar los datos mediante el algoritmo de clusters jerárquico. Para esto el primer paso es estandarizar las muestras para evitar errores de clasificación. Los resultados arrojados los podemos apreciar en la gráfica 1, en estos se ve como todos los datos son semejantes, esto se debe al alto número de variables y la poca correlación que existe entre ellas imposibilitando agrupar los datos en clusters.

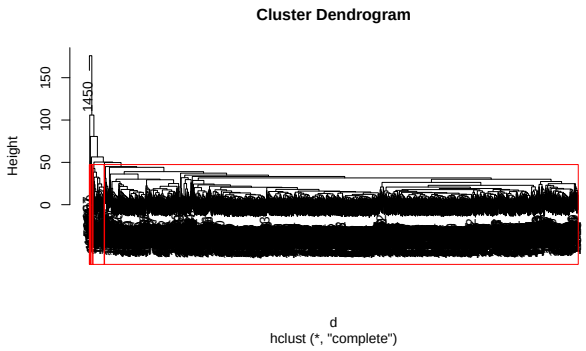

Figure 1: clusters jerárquico de los datos con todas las variables

Para aplicar el algoritmo de K-Means debemos conocer en cuantos clusters intentaremos dividir la muestra, para esto hallamos el error que se obtiene al intentar particionar los datos en distintos números de clusters y tomamos el que mejor relación tenga. Como podemos ver en la grafica 2 para una cantidad de clusters mayor que 8 el error disminuye pero en menor escala que como lo hacía antes, y como el objetivo es tener la menor cantidad de clusters posible, tomamos 8 ya que posee un menor error y una menor cantidad de clusters.

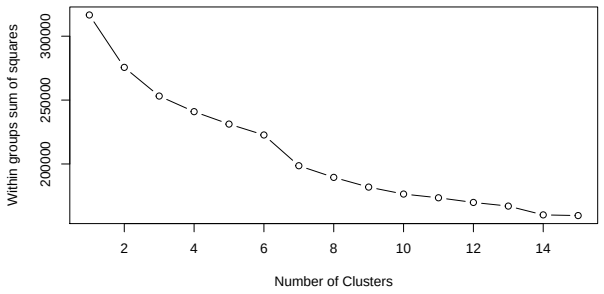

Figure 2: Error de los datos por cada cantidad de clusters

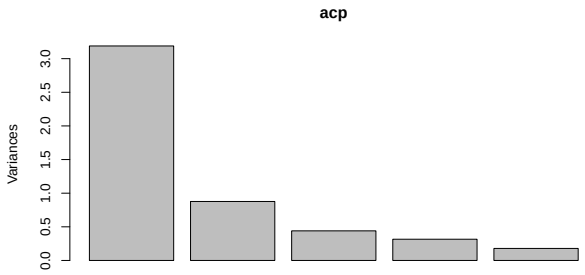

Figure 3: Proporción acumulativa de componentes principales

Ahora con la cantidad de clusters definida podemos aplicar el algoritmo quedando una distribución que tiene una medida de similitud de los elementos dentro de cada cluster de 40.6%, la cual es baja debido a la cantidad de variables que poseemos sin ninguna correlación entre sí.

Ahora intentaremos disminuir la dimensión de los datos tomando solo un conjunto de las variables que posee el conjunto de muestras, en particular son las que ya habíamos tomado en los anteriores acercamientos.

Veamos nuevamente la correlación entre las variables, pero esta vez de las variables seleccionadas.

```
                    m PP PU V n
medIncome           1
PctPopUnderPov      , 1
PctUnemployed       , ,  1
ViolentCrimesPerPop . .  .  1
nonViolPerPop       . .  .  , 1
```

A pesar de ver un número menor de espacios en blanco, se puede ver como todo son . y , lo que hace que igual no sea un matriz altamente correlacionada, veamos que resultados arroja aplicar el algoritmo de ACP, para esto igualmente estandarizamos los datos. Obtenemos que con solo tomar las dos primeras componentes principales estas obtendriamos una proporción acumulativa de 0.8131 con lo que explicaríamos el 81.31% de la variación de los datos, este por ciento es mayor que 70% por lo que es aceptable, esto se puede apreciar mejor en la grafica 3. Si nos guiásemos por el criterio de Kaiser tenemos que solo la primera componente posee valor propio mayor que 1.

Veamos qué variables son importantes para la componente 1. Podemos tomar el mayor valor propio absoluto que es 0.5021, teniendo este podemos afirmar que todos los valores propios cuyo módulo esté por encima de 0.2510 en la columna de PC1 son las variables que conforman esta componente. Por tanto PC1 está caracterizada por personas con bajos ingresos, gran porciento de personas por debajo del nivel de pobreza, alto porciento de personas en edad laboral desempleados, alto número de crímenes violentos y alto número de crímenes no violentos.

Siguiendo el análisis que se realizó teniendo en cuenta todas las variables, podemos aplicar el algoritmo de clústers jerárquicos pero con este obtenemos un resultado similar al anterior en donde se obtiene un clusters que posee a la mayoría de los datos y el resto solo posee una pequeña proporción de las muestras.

Ahora veamos para este conjunto de variables cuál sería el mejor número de cluster a formar con el algorítmo de K-Means, realizando un análisis similar podemos ver en la gráfica 4 como para estas variables el mejor número de clusters a tomar sería 4. Con esta distribución obtenemos una medida de similitud de los elementos dentro de cada cluster de 69.5%, la cual es más alta que el anterior análisis a pesar de poseer un menor número de clusters.

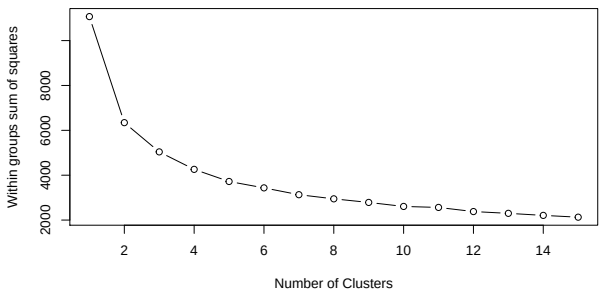

Figure 4: Error de los datos por cada cantidad de clusters

Podemos ver la distribución de clusters para las muestras con respecto a las variables seleccionadas en la figura 5.

5. ANOVA

El análisis de varianza (ANOVA) permite determinar si las diferencias entre grupos de datos son significativas estadísiticamente. En el caso particular de el set de datos utilizado en este proyecto, se tiene que las vari-

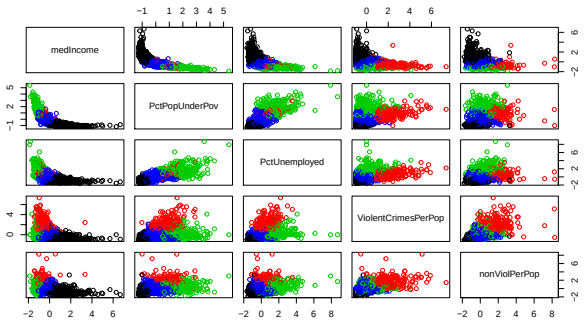

Figure 5: Datos representados con colores diferenciando su pertenencia a cada cluster

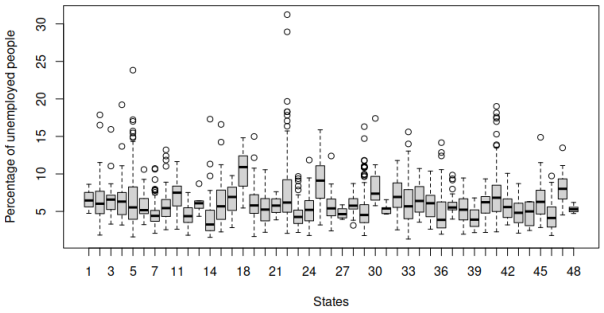

Figure 6: Bloxplot de la tasa de desempleo para cada estado

```
> unemployed.anova <- aov(unemployed~states_map, data=df_unemp)
> summary(unemployed.anova)
             Df Sum Sq Mean Sq F value Pr(>F)
states_map    1     13  13.459   1.604  0.205
Residuals  2210  18545   8.391
```

Figure 7: Resultados de aplicar la técnica ANOVA sobre los datos

ables se estudian sobre las diferentes comunidades de los estados de los Estados Unidos. Por ello se puede llevar a cabo un estudio de dichas variables y su variación con respecto a los diferentes estados. Previamente fueron seleccionadas algunas variables para su estudio, las mismas se estarán utilizando en estos casos.

Note que la cantidad de comunidades por cada estado varía en dependencia del mismo, por tanto en primer lugar se debió hacer un filtrado de la base de datos, eliminando aquellos estados que solo contaban con una observación (o lo que es lo mismo, que solo contaban con una comunidad como muestra). Igualmente, en el caso de las variables *ViolentCrimesPerPop* y *nonViolPerPop*, ocurría que habían datos faltantes; no obstante como la cantidad de ellos era muy pequeña respecto al resto de los datos, pudieron ser obviadas estas filas. Además, dado que el Estado al que pertenece una comunidad es un dato categórico, se lleva a cabo el mapeo del vector de Estados, tal que a cada uno se le asigna un número. El código presente en el archivo anova.R contiene la implementación de estos métodos para cada variable.

Tomemos el caso de la variable *PctUnemployed*, que representa la cantidad de personas desempleadas por cada 100000 habitantes en las comunidades, y que a su vez es de gran importancia en le análisis de criminalidad de las poblaciones dado que se encuentra estrechamente relacionada con los ingresos medios de las familias, el nivel de pobreza, etc. En la figura 6 se puede observar la distribución de los datos según los Estados.

Se lleva a cabo el ANOVA obteniendo los resultados que se muestran en la figura 7:

Como se puede observar el $valor - p = 0.205$ y este valor es mayor que la significación prefijada ($\alpha = 0.05$), por tanto H_0 no puede ser rechazada lo q implica que los valores de varianza de los Estados con repecto al desempleo se encuentran alrededor del mismo valor. Tras este cómputo se procede a hallar los residuos para el cálculo delos supuestos.

Al llevar a cabo la prueba de Shapiro-Wilk, se obtiene el siguiente resultado (8) Note que el $p - value < 0.05$, que es la significación prefijada, lo que implica que es significativo y por tanto se rechaza H_0. Al ocurrir esto, se puede afirmar que los residuos no presentan una distribución normal. Al incumplirse uno de ellos no es necesario verificar los otros supuestos, puesto que el proyecto puede ser rechazado.

Análogamente se realiza este análisis con el resto de las variables que se deseaba estudiar. En cada caso se obtiene que no se cumplen los supuestos y por tanto se rechaza el modelo.

6. Conclusiones

Se llevó a cabo el estudio de los datos a través de los ejercicios propuestos, utilizando las técnicas de regresión, reducción de dimensión y ANOVA y posteriormente se desarrolló la interpretación de los resultados obtenidos.

En el caso de ANOVA, lamentablemente se incumplieron los supuestos en cada una de los modelos deter-

```
> shapiro.test(res)

        Shapiro-Wilk normality test

data:  res
W = 0.81112, p-value < 2.2e-16
```

Figure 8: Resultados de aplicar la prueba Shapiro-Wilks sobre los residuos

minados por las variables seleccionadas, lo que permite afirmar que no se pudieron aplicar las técnicas de anova sobre estos datos. Igualmente ocurrió en el caso de los modelos de regresión lineal, donde los supuestos se incumplieron, por lo que se reafirma que dichos modelos no fueron buenos. Este análisis de los datos indica que , por lo menos en el caso de las variables seleccionadas, se encuentran sesgados y poco correlacionados.

Por último, en el caso de la aplicación de técnicas para reducción de la dimensión se vuelve a observar la poca correlación existente entre los datos. En el caso de las variables seleccionadas, elegir un número menor para realizar el algoritmo de $K - Means$ proporcionó mejores resultados.

7. Contribución de los integrantes

Esta segunda fase del proyecto requiere la aplicación de tres técnicas diferentes para el tratamiento de los datos, que evalúan distintas habilidades impartidas en clase. Se siguió como estrategia, al igual que en la primera fase, la lectura y debate iniciales de los problemas por todos los integrantes, de modo que quedara determinado el la ruta para la solución de cada uno. Una vez terminado este análisis se decidió que cada integrante se responsabilizara con la realización del código y la parte correspondiente del informe de una de las técnicas; tal distribución se obtuvo de forma aleatoria, quedando que:

- Regresión lineal: Osmany Pérez
- Reducción de dimensión: Enrique Martínez
- ANOVA: Carmen I. Cabrera

Los archivos correspondientes al proyecto están almacenados en un repositorio de github, donde podrá corroborar que todos los participantes contribuyeron a su realización. Puede acceder a dicho repositorio a través del siguiente link