

# Segundo Proyecto Estadística. Curso 2019-2020

**Daniel Alberto García Pérez**  
Grupo C412

[D.GARCIA@ESTUDIANTES.MATCOM.UH.CU](mailto:D.GARCIA@ESTUDIANTES.MATCOM.UH.CU)

**Leonel Alejandro García López**  
Grupo C412

[L.GARCIA3@ESTUDIANTES.MATCOM.UH.CU](mailto:L.GARCIA3@ESTUDIANTES.MATCOM.UH.CU)

**Roberto Marti Cedeño**  
Grupo C412

[R.MARTI@ESTUDIANTES.MATCOM.UH.CU](mailto:R.MARTI@ESTUDIANTES.MATCOM.UH.CU)

## Tutor(es):

Msc. Dalia Diaz Sistachs, *Facultad de Matemática y Computación, Universidad de La Habana*

**Tema:** Estadística, Técnicas de Clasificación, Regresión, Anova.

## 1. Introducción

El siguiente informe corresponde al trabajo de los autores como parte de la investigación realizada sobre los datos asignados en su segundo proyecto de la asignatura.

Los datos asignados, responden a un estudio realizado sobre las respuestas correspondientes a un sensor de gases en una ciudad italiana (Data/AirQualityUCI.csv). Se tuvieron en cuenta las respuestas de los distintos terminales del sensor, así como las concentraciones de gases existentes en el ambiente.

Por razones desconocidas, existen observaciones incompletas de cada una de las variables presentes en la recopilación, por lo que se hace necesario modificar las mismas para poder realizar un estudio acorde con los requerimientos de cada uno de los métodos a emplear.

### 1.1 Descripción Inicial

Para la descripción de las características generales de los datos, se empleó el método *skim* presente en la biblioteca *skimr* de R, el cual brinda, entre sus valores principales, la cantidad de datos faltantes, así como los estadísticos descriptivos de cada una de las variables de la muestra.

```
-- Data Summary -----
Name                Values
Number of rows      9357
Number of columns    15

Column type frequency:
factor              2
numeric            13

Group variables      None

-- variable type: factor -----
# A tibble: 2 x 6
  skim_variable n_missing complete_rate ordered n_unique top_counts
  <chr>          <int>         <dbl> <dbl>      <int> <chr>
1 Date           0           1 FALSE   391 01/: 24, 01/: 24, 01/: 24
2 Time           0           1 FALSE   24 00.: 390, 01.: 390, 02.: 390, 03.: 390
```

Figure 1: Descripción de la fuente de datos mediante la función skim (Parte 1).

Como se puede apreciar de la primera mitad de los datos obtenidos de la función (Figura 1), el set de datos se compone por 9357 observaciones de 15 variables. Estas se componen por 2 de tipo factor y 13 numéricas. Es importante destacar que dado que las dos variables de tipo factor, la fecha y la hora de las mediciones, se descartaron por el equipo para el análisis dado que, las mediciones se realizaron exclusivamente durante tres meses, por lo cual, no se cuenta con información suficiente para caracterizar el resto de los resultados a partir del tiempo.

```
-- variable type: numeric -----
# A tibble: 13 x 11
  skim_variable n_missing complete_rate mean      sd      p0      p25      p50      p75     p100 hist
  <chr>          <int>         <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
1 CO.GT.         1683      0.820    2.15    1.45    0.1     1.1     1.8     2.9    11.9  [ ]
2 PT08.S1.CO.    366      0.961  1100.    217.    647    937    1063    1231    2040  [ ]
3 NHHC.GT.       8443      0.0977   219.    204.    7      67     150    297    1189  [ ]
4 C6H6.GT.       366      0.961   10.1     7.45    0.1     4.4     8.2     14     63.7  [ ]
5 PT08.S2.NHHC.  366      0.961   939.    267.    383    734    909    1116    2214  [ ]
6 NOX.GT.        1639      0.825   247.    213.    2      98    180    326    1479  [ ]
7 PT08.S3.NOX.   366      0.961   835.    257.    322    658    806    970.    2683  [ ]
8 NO2.GT.        1642      0.825   113.    48.4    2      78    109    142    340   [ ]
9 PT08.S4.NO2.   366      0.961  1456.    346.    551   1227   1463   1674    2775  [ ]
10 PT08.S5.O3.   366      0.961  1023.    398.    221   732.    963   1274.   2523  [ ]
11 T              366      0.961   18.3     8.83   -1.9   11.8   17.8   24.4    44.6  [ ]
12 RH             366      0.961   49.2    17.3    9.2   35.8   49.6   62.5    88.7  [ ]
13 AH             366      0.961    1.03    0.404   0.185  0.737  0.995  1.31    2.23  [ ]
```

Figure 2: Descripción de la fuente de datos mediante la función skim (Parte 2).

Centrando el análisis en la segunda parte de la respuesta obtenida de *skim* (Figura 2), resalta la cantidad de observaciones faltantes en cada una de las variables, que varían desde 366, hasta 8443. Las variables presentes, son de forma general de 3 tipos, las respuestas de los sensores a determinados compuestos del aire, la concentración de los compuestos presente, y variables generales del ambiente, temperatura, humedad relativa y absoluta.

Para un mejor empleo de los datos, se completaron los datos faltantes con la media de cada una de las variables descritas en los datos.

## 2. Regresión y ACP

El primer análisis a realizar sobre los datos fue la regresión, para ello se tuvo en cuenta la correlación existente entre cada una de las variables numéricas para corroborar si sería de utilidad realizarla.

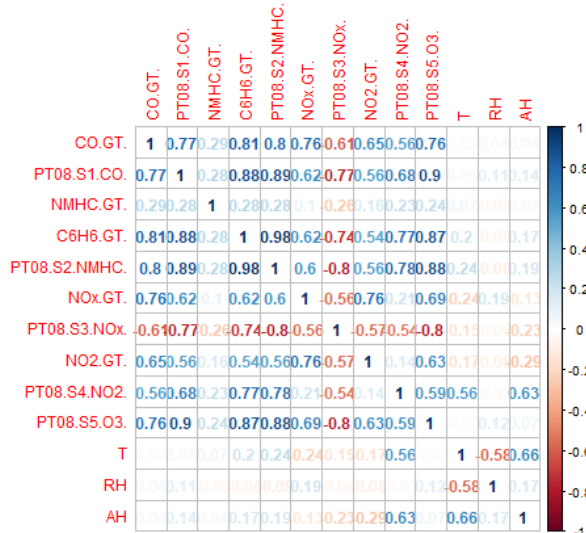


Figure 3: Correlación entre las variables.

La primera característica presente en los datos que se observa a partir de su correlación (Figura 3) es que los datos de los compuestos están altamente correlacionados, por lo que la realización de una regresión lineal sobre los mismos incurriría en el problema de la multicolinealidad.

Como estrategia de solución al problema anterior, se decidió agrupar los datos por sus componentes principales, para así, reducir los datos de ser posible y poder dar una interpretación mas clara a la regresión. Como no se dispone de interés especial por alguna variable se decidió empelar la variable de la respuesta del sensor número 4 relacionado con el dióxido de nitrógeno ( $PT08.S4.NO_2$ ) debido a que presenta pocos datos faltantes y es la mas correlacionada con los datos disponibles.

### 2.1 ACP

A continuación (Figura 4) se encuentran todas las posibles componentes resultantes del desglose (Buscar como se escribe) de los datos.

Importance of components:								
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	2.4959	1.4415	1.1395	0.9612	0.71574	0.63999	0.40913	0.3763
Proportion of Variance	0.5191	0.1732	0.1082	0.0770	0.04269	0.03413	0.01395	0.0118
Cumulative Proportion	0.5191	0.6923	0.8005	0.8775	0.92018	0.95431	0.96826	0.9801
	PC9	PC10	PC11	PC12				
Standard deviation	0.32477	0.29453	0.19141	0.10205				
Proportion of Variance	0.00879	0.00723	0.00305	0.00087				
Cumulative Proportion	0.98885	0.99608	0.99913	1.00000				

Figure 4: Posibles componentes principales.

Se tuvieron en cuenta dos criterios tomados de la literatura para la selección de las componentes principales, un criterio de *porcentaje* que debería superar como mínimo el 70% de los datos y el criterio de *Kaiser*. Como podemos observar en las componentes (Figura 4) con las dos primeras ya se cumple el criterio de mas del 70% de los datos, pero para cumplir también con *Kaiser* se extendieron las componentes hasta la tercera.

#### 2.1.1 DESCRIPCIÓN DE LAS COMPONENTES

Para poder describir detalladamente las características de cada una de las tres componentes en las que se aglomeran los datos de estudio se analizó su matriz de valores propios. (Figura 5)

	PC1	PC2	PC3
CO. GT.	-0.35347648	-0.04638393	0.03504330
PT08. S1. CO.	-0.37073311	0.02593933	-0.10493234
NMHC. GT.	-0.12883560	0.09317803	0.08870589
C6H6. GT.	-0.37361310	0.12266320	0.01402562
PT08. S2. NMHC.	-0.37801424	0.14739559	0.02693871
NOx. GT.	-0.31179657	-0.26951354	0.01283753
PT08. S3. NOx.	0.33774833	-0.09221096	0.09989306
NO2. GT.	-0.29045629	-0.24388495	0.30649399
PT08. S5. O3.	-0.37473906	-0.03834792	-0.08045593
T	-0.02533438	0.66090296	0.14332237
RH	-0.01688014	-0.32692259	-0.75967656
AH	-0.03897675	0.51571297	-0.52035094

Figure 5: Valores propios de las componentes principales.

La primera componente se caracteriza por bajos valores de la concentración de todos los compuestos presentes en el estudio con excepción los hidrocarburos no metánicos ( $NMHC.GT$ ), así como bajos valores de respuesta de todos los sensores, exceptuando a la variable dependiente. Esta componente, la mas numerosa de las analizadas describe el comportamiento mas común de los datos, este resultado puede estar determinado por que las muestras se obtuvieron en una sola zona de la ciudad, en una sola ciudad o en un intervalo de tiempo donde no varían mucho.

La segunda componente se caracteriza por valores altos de la temperatura y humedad relativa. Esta componente describe las situaciones de las horas cercanas al mediodía donde la temperatura es mas elevada.

La última componente se caracteriza por altos valores de la humedad relativa y absoluta. Dada la presencia de altos valores de humedad tanto en la 2da como en la 3ra componente se puede llegar a la conclusión que algunas de las mediciones se encontraron en temporada de lluvias u ocurrió algún evento climatológico.

### 2.2 Regresión

Posterior a la definición de las componentes se dispuso la creación de un modelo de regresión múltiple, donde la variable dependiente se tomó como la respuesta del cuarto sensor a la concentración de  $NO_2$ , y como variables independientes las 3 componentes resultantes del

ACP. Se comprobó una vez mas la utilidad de realizar la regresión mediante los gráficos de dispersión y correlación (Figuras 6 y 7).

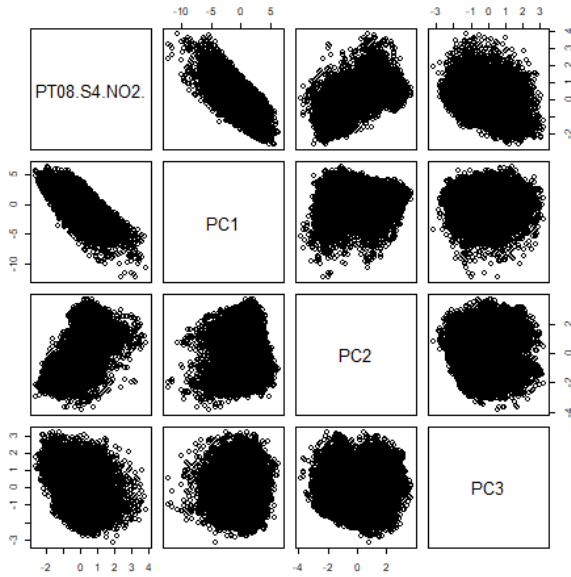


Figure 6: Gráfico de dispersión.

Del gráfico de dispersión (Figura 6), podemos percatarnos de la correlación inversa de la variable dependiente con la primera componente, así como su relación débil pero lineal con la segunda. Ambos datos se esclarecen con la gráfica de correlación (Figura 7).

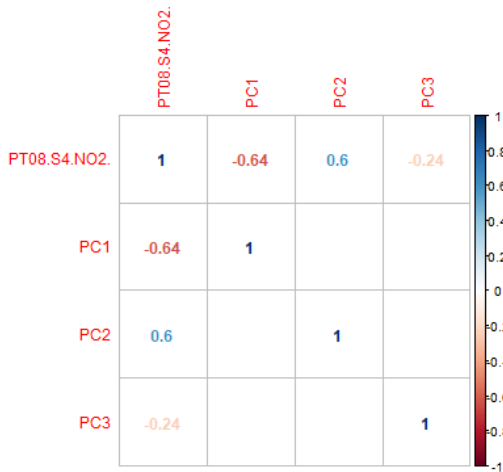


Figure 7: Gráfico de correlación.

Otro detalle significativo se deduce de los resultados de la clasificación en componentes principales, dado que brindan una segmentación en variables independientes, como se muestra en la figura 7. Finalmente

antes de la realización de la regresión se tomó como consenso la admisión de los valores de las correlaciones de la primera y segunda componentes con la variable dependiente como lineales.

```
Call:
lm(formula = formula, data = as.data.frame(dataset))

Residuals:
    Min       1Q   Median       3Q      Max
-2.0099 -0.3096 -0.1014  0.3036  1.9046

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e-16  4.205e-03     0.00    1
PC1          -2.580e-01  1.685e-03  -153.09 <2e-16 ***
PC2           4.165e-01  2.917e-03   142.76 <2e-16 ***
PC3          -2.142e-01  3.691e-03   -58.03 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4068 on 9353 degrees of freedom
Multiple R-squared:  0.8346, Adjusted R-squared:  0.8345
F-statistic: 1.573e+04 on 3 and 9353 DF, p-value: < 2.2e-16
```

Figure 8: Resumen de la regresión.

Los resultados obtenidos del modelo se reflejan en la figura 8. De los mismos se deduce la siguiente ecuación para determinar el valor de la respuesta del sensor a la concentración de dióxido de nitrógeno.

$$PT08.S4.NO_2 = -0.26 * PC1 + 0.42 * PC2 - 0.22 * PC3$$

Uno de los factores a tener en cuenta es la estandarización de los datos durante el proceso de clasificación, por lo que su efecto se ve reflejado en la ausencia del término independiente en la ecuación de regresión.

En el modelo propuesto por cada unidad de decremento de todas las variables presentes en la componente 1 (Ver sección 2.1.1), el valor de respuesta del sensor al dióxido de nitrógeno disminuye en 0.26 unidades aproximadamente. Por cada unidad de incremento en la temperatura y la humedad absoluta, se incrementa la respuesta del sensor en aproximadamente 0.42 y finalmente por cada unidad de incremento de la humedad relativa y absoluta se decrementa la respuesta en 0.22 unidades aproximadamente.

La precisión del modelo medida en términos del valor de  $R - ajustado$  es de 0.8343 lo cual es bastante alto tomando en consideración los datos faltantes y el desprecio de datos resultante del ACP. El  $p - valor$  de la prueba de  $F - statistic$  es menor que 0.05 por lo que podemos asegurar que nuestro modelo produce resultados. El error residual es de 0.4. Todas las variables independientes son de importancia para la estimación de la variable dependiente.

## 2.2.1 ANÁLISIS DE LOS SUPUESTOS DE LA REGRESIÓN

1. Las variables independientes no están correlacionadas.
2. La media y la suma de los errores es cero.
3. Los errores tienen distribución normal.
4. Los errores son independientes.

## 5. La varianza de los errores es constante.

El primer requisito de los supuestos del modelo se cumple al emplear como variables independientes el resultado de aplicar el *ACP*. (Ver gráfico 7).

Se obtuvo del modelo que la suma de los errores es  $4.3e - 13$  y la media de los mismos es  $4.6e - 17$  por lo que podemos asegurar que la media y la suma de los errores es 0.

```

Durbin-watson test

data: regression
DW = 0.152, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0

```

Figure 9: Prueba de Durwin-Watson.

Como se puede observar del resultado de la prueba de independencia de Durwin-Watson (Figura 9), el  $p$ -valor  $<< 0.05$  por lo que podemos rechazar la hipótesis nula y los errores son dependientes. Esto incumple con los supuestos del modelo por lo que se termina su análisis y se descarta su empleo.

En la carpeta de imágenes adjunta al proyecto se pueden visualizar los resultados obtenidos de la normalidad de los errores y la homocedasticidad.

## 3. Cúlster, Kmeans y Árbol de Decisión

Para el análisis de tipo clúster se tomaron en cuenta los resultados obtenidos del *ACP*, por lo que el número de particiones del clúster jerárquico se fijó a 3, pero también se valoró la inclusión de 2 clústers como se puede apreciar en el dendrograma (Figura 10)

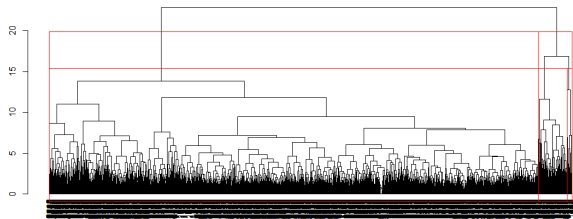


Figure 10: Dendrograma del clúster jerárquico con 2 y 3 particiones de los datos.

Uno de los detalles que más denotan de la gráfica anterior recae en la gran acumulación de datos en la primera partición, posiblemente se deba a que los datos son tomados de una misma zona y no variaron mucho durante los tres meses del estudio.

Tomando los resultados anteriores, se dispuso de la ejecución del algoritmo *Kmeans* cuyo resultado se puede observar en la figura 11. Primero se intentó una aproximación de dos particiones, pero se obtuvo una similitud entre componentes de un 33%, por lo que se

```

K-means clustering with 3 clusters of sizes 2144, 3913, 3300

Cluster means:
  CO2.GT  PM10.11.CO  PM10.11.CO2  PM10.11.CO3  PM10.11.CO4  PM10.11.CO5  PM10.11.CO6  PM10.11.CO7  PM10.11.CO8  PM10.11.CO9  PM10.11.CO10  PM10.11.CO11  PM10.11.CO12  PM10.11.CO13  PM10.11.CO14  PM10.11.CO15  PM10.11.CO16  PM10.11.CO17  PM10.11.CO18  PM10.11.CO19  PM10.11.CO20  PM10.11.CO21  PM10.11.CO22  PM10.11.CO23  PM10.11.CO24  PM10.11.CO25  PM10.11.CO26  PM10.11.CO27  PM10.11.CO28  PM10.11.CO29  PM10.11.CO30  PM10.11.CO31  PM10.11.CO32  PM10.11.CO33  PM10.11.CO34  PM10.11.CO35  PM10.11.CO36  PM10.11.CO37  PM10.11.CO38  PM10.11.CO39  PM10.11.CO40  PM10.11.CO41  PM10.11.CO42  PM10.11.CO43  PM10.11.CO44  PM10.11.CO45  PM10.11.CO46  PM10.11.CO47  PM10.11.CO48  PM10.11.CO49  PM10.11.CO50  PM10.11.CO51  PM10.11.CO52  PM10.11.CO53  PM10.11.CO54  PM10.11.CO55  PM10.11.CO56  PM10.11.CO57  PM10.11.CO58  PM10.11.CO59  PM10.11.CO60  PM10.11.CO61  PM10.11.CO62  PM10.11.CO63  PM10.11.CO64  PM10.11.CO65  PM10.11.CO66  PM10.11.CO67  PM10.11.CO68  PM10.11.CO69  PM10.11.CO70  PM10.11.CO71  PM10.11.CO72  PM10.11.CO73  PM10.11.CO74  PM10.11.CO75  PM10.11.CO76  PM10.11.CO77  PM10.11.CO78  PM10.11.CO79  PM10.11.CO80  PM10.11.CO81  PM10.11.CO82  PM10.11.CO83  PM10.11.CO84  PM10.11.CO85  PM10.11.CO86  PM10.11.CO87  PM10.11.CO88  PM10.11.CO89  PM10.11.CO90  PM10.11.CO91  PM10.11.CO92  PM10.11.CO93  PM10.11.CO94  PM10.11.CO95  PM10.11.CO96  PM10.11.CO97  PM10.11.CO98  PM10.11.CO99  PM10.11.CO100  PM10.11.CO101  PM10.11.CO102  PM10.11.CO103  PM10.11.CO104  PM10.11.CO105  PM10.11.CO106  PM10.11.CO107  PM10.11.CO108  PM10.11.CO109  PM10.11.CO110  PM10.11.CO111  PM10.11.CO112  PM10.11.CO113  PM10.11.CO114  PM10.11.CO115  PM10.11.CO116  PM10.11.CO117  PM10.11.CO118  PM10.11.CO119  PM10.11.CO120  PM10.11.CO121  PM10.11.CO122  PM10.11.CO123  PM10.11.CO124  PM10.11.CO125  PM10.11.CO126  PM10.11.CO127  PM10.11.CO128  PM10.11.CO129  PM10.11.CO130  PM10.11.CO131  PM10.11.CO132  PM10.11.CO133  PM10.11.CO134  PM10.11.CO135  PM10.11.CO136  PM10.11.CO137  PM10.11.CO138  PM10.11.CO139  PM10.11.CO140  PM10.11.CO141  PM10.11.CO142  PM10.11.CO143  PM10.11.CO144  PM10.11.CO145  PM10.11.CO146  PM10.11.CO147  PM10.11.CO148  PM10.11.CO149  PM10.11.CO150  PM10.11.CO151  PM10.11.CO152  PM10.11.CO153  PM10.11.CO154  PM10.11.CO155  PM10.11.CO156  PM10.11.CO157  PM10.11.CO158  PM10.11.CO159  PM10.11.CO160  PM10.11.CO161  PM10.11.CO162  PM10.11.CO163  PM10.11.CO164  PM10.11.CO165  PM10.11.CO166  PM10.11.CO167  PM10.11.CO168  PM10.11.CO169  PM10.11.CO170  PM10.11.CO171  PM10.11.CO172  PM10.11.CO173  PM10.11.CO174  PM10.11.CO175  PM10.11.CO176  PM10.11.CO177  PM10.11.CO178  PM10.11.CO179  PM10.11.CO180  PM10.11.CO181  PM10.11.CO182  PM10.11.CO183  PM10.11.CO184  PM10.11.CO185  PM10.11.CO186  PM10.11.CO187  PM10.11.CO188  PM10.11.CO189  PM10.11.CO190  PM10.11.CO191  PM10.11.CO192  PM10.11.CO193  PM10.11.CO194  PM10.11.CO195  PM10.11.CO196  PM10.11.CO197  PM10.11.CO198  PM10.11.CO199  PM10.11.CO200  PM10.11.CO201  PM10.11.CO202  PM10.11.CO203  PM10.11.CO204  PM10.11.CO205  PM10.11.CO206  PM10.11.CO207  PM10.11.CO208  PM10.11.CO209  PM10.11.CO210  PM10.11.CO211  PM10.11.CO212  PM10.11.CO213  PM10.11.CO214  PM10.11.CO215  PM10.11.CO216  PM10.11.CO217  PM10.11.CO218  PM10.11.CO219  PM10.11.CO220  PM10.11.CO221  PM10.11.CO222  PM10.11.CO223  PM10.11.CO224  PM10.11.CO225  PM10.11.CO226  PM10.11.CO227  PM10.11.CO228  PM10.11.CO229  PM10.11.CO230  PM10.11.CO231  PM10.11.CO232  PM10.11.CO233  PM10.11.CO234  PM10.11.CO235  PM10.11.CO236  PM10.11.CO237  PM10.11.CO238  PM10.11.CO239  PM10.11.CO240  PM10.11.CO241  PM10.11.CO242  PM10.11.CO243  PM10.11.CO244  PM10.11.CO245  PM10.11.CO246  PM10.11.CO247  PM10.11.CO248  PM10.11.CO249  PM10.11.CO250  PM10.11.CO251  PM10.11.CO252  PM10.11.CO253  PM10.11.CO254  PM10.11.CO255  PM10.11.CO256  PM10.11.CO257  PM10.11.CO258  PM10.11.CO259  PM10.11.CO260  PM10.11.CO261  PM10.11.CO262  PM10.11.CO263  PM10.11.CO264  PM10.11.CO265  PM10.11.CO266  PM10.11.CO267  PM10.11.CO268  PM10.11.CO269  PM10.11.CO270  PM10.11.CO271  PM10.11.CO272  PM10.11.CO273  PM10.11.CO274  PM10.11.CO275  PM10.11.CO276  PM10.11.CO277  PM10.11.CO278  PM10.11.CO279  PM10.11.CO280  PM10.11.CO281  PM10.11.CO282  PM10.11.CO283  PM10.11.CO284  PM10.11.CO285  PM10.11.CO286  PM10.11.CO287  PM10.11.CO288  PM10.11.CO289  PM10.11.CO290  PM10.11.CO291  PM10.11.CO292  PM10.11.CO293  PM10.11.CO294  PM10.11.CO295  PM10.11.CO296  PM10.11.CO297  PM10.11.CO298  PM10.11.CO299  PM10.11.CO300  PM10.11.CO301  PM10.11.CO302  PM10.11.CO303  PM10.11.CO304  PM10.11.CO305  PM10.11.CO306  PM10.11.CO307  PM10.11.CO308  PM10.11.CO309  PM10.11.CO310  PM10.11.CO311  PM10.11.CO312  PM10.11.CO313  PM10.11.CO314  PM10.11.CO315  PM10.11.CO316  PM10.11.CO317  PM10.11.CO318  PM10.11.CO319  PM10.11.CO320  PM10.11.CO321  PM10.11.CO322  PM10.11.CO323  PM10.11.CO324  PM10.11.CO325  PM10.11.CO326  PM10.11.CO327  PM10.11.CO328  PM10.11.CO329  PM10.11.CO330  PM10.11.CO331  PM10.11.CO332  PM10.11.CO333  PM10.11.CO334  PM10.11.CO335  PM10.11.CO336  PM10.11.CO337  PM10.11.CO338  PM10.11.CO339  PM10.11.CO340  PM10.11.CO341  PM10.11.CO342  PM10.11.CO343  PM10.11.CO344  PM10.11.CO345  PM10.11.CO346  PM10.11.CO347  PM10.11.CO348  PM10.11.CO349  PM10.11.CO350  PM10.11.CO351  PM10.11.CO352  PM10.11.CO353  PM10.11.CO354  PM10.11.CO355  PM10.11.CO356  PM10.11.CO357  PM10.11.CO358  PM10.11.CO359  PM10.11.CO360  PM10.11.CO361  PM10.11.CO362  PM10.11.CO363  PM10.11.CO364  PM10.11.CO365  PM10.11.CO366  PM10.11.CO367  PM10.11.CO368  PM10.11.CO369  PM10.11.CO370  PM10.11.CO371  PM10.11.CO372  PM10.11.CO373  PM10.11.CO374  PM10.11.CO375  PM10.11.CO376  PM10.11.CO377  PM10.11.CO378  PM10.11.CO379  PM10.11.CO380  PM10.11.CO381  PM10.11.CO382  PM10.11.CO383  PM10.11.CO384  PM10.11.CO385  PM10.11.CO386  PM10.11.CO387  PM10.11.CO388  PM10.11.CO389  PM10.11.CO390  PM10.11.CO391  PM10.11.CO392  PM10.11.CO393  PM10.11.CO394  PM10.11.CO395  PM10.11.CO396  PM10.11.CO397  PM10.11.CO398  PM10.11.CO399  PM10.11.CO400  PM10.11.CO401  PM10.11.CO402  PM10.11.CO403  PM10.11.CO404  PM10.11.CO405  PM10.11.CO406  PM10.11.CO407  PM10.11.CO408  PM10.11.CO409  PM10.11.CO410  PM10.11.CO411  PM10.11.CO412  PM10.11.CO413  PM10.11.CO414  PM10.11.CO415  PM10.11.CO416  PM10.11.CO417  PM10.11.CO418  PM10.11.CO419  PM10.11.CO420  PM10.11.CO421  PM10.11.CO422  PM10.11.CO423  PM10.11.CO424  PM10.11.CO425  PM10.11.CO426  PM10.11.CO427  PM10.11.CO428  PM10.11.CO429  PM10.11.CO430  PM10.11.CO431  PM10.11.CO432  PM10.11.CO433  PM10.11.CO434  PM10.11.CO435  PM10.11.CO436  PM10.11.CO437  PM10.11.CO438  PM10.11.CO439  PM10.11.CO440  PM10.11.CO441  PM10.11.CO442  PM10.11.CO443  PM10.11.CO444  PM10.11.CO445  PM10.11.CO446  PM10.11.CO447  PM10.11.CO448  PM10.11.CO449  PM10.11.CO450  PM10.11.CO451  PM10.11.CO452  PM10.11.CO453  PM10.11.CO454  PM10.11.CO455  PM10.11.CO456  PM10.11.CO457  PM10.11.CO458  PM10.11.CO459  PM10.11.CO460  PM10.11.CO461  PM10.11.CO462  PM10.11.CO463  PM10.11.CO464  PM10.11.CO465  PM10.11.CO466  PM10.11.CO467  PM10.11.CO468  PM10.11.CO469  PM10.11.CO470  PM10.11.CO471  PM10.11.CO472  PM10.11.CO473  PM10.11.CO474  PM10.11.CO475  PM10.11.CO476  PM10.11.CO477  PM10.11.CO478  PM10.11.CO479  PM10.11.CO480  PM10.11.CO481  PM10.11.CO482  PM10.11.CO483  PM10.11.CO484  PM10.11.CO485  PM10.11.CO486  PM10.11.CO487  PM10.11.CO488  PM10.11.CO489  PM10.11.CO490  PM10.11.CO491  PM10.11.CO492  PM10.11.CO493  PM10.11.CO494  PM10.11.CO495  PM10.11.CO496  PM10.11.CO497  PM10.11.CO498  PM10.11.CO499  PM10.11.CO500  PM10.11.CO501  PM10.11.CO502  PM10.11.CO503  PM10.11.CO504  PM10.11.CO505  PM10.11.CO506  PM10.11.CO507  PM10.11.CO508  PM10.11.CO509  PM10.11.CO510  PM10.11.CO511  PM10.11.CO512  PM10.11.CO513  PM10.11.CO514  PM10.11.CO515  PM10.11.CO516  PM10.11.CO517  PM10.11.CO518  PM10.11.CO519  PM10.11.CO520  PM10.11.CO521  PM10.11.CO522  PM10.11.CO523  PM10.11.CO524  PM10.11.CO525  PM10.11.CO526  PM10.11.CO527  PM10.11.CO528  PM10.11.CO529  PM10.11.CO530  PM10.11.CO531  PM10.11.CO532  PM10.11.CO533  PM10.11.CO534  PM10.11.CO535  PM10.11.CO536  PM10.11.CO537  PM10.11.CO538  PM10.11.CO539  PM10.11.CO540  PM10.11.CO541  PM10.11.CO542  PM10.11.CO543  PM10.11.CO544  PM10.11.CO545  PM10.11.CO546  PM10.11.CO547  PM10.11.CO548  PM10.11.CO549  PM10.11.CO550  PM10.11.CO551  PM10.11.CO552  PM10.11.CO553  PM10.11.CO554  PM10.11.CO555  PM10.11.CO556  PM10.11.CO557  PM10.11.CO558  PM10.11.CO559  PM10.11.CO560  PM10.11.CO561  PM10.11.CO562  PM10.11.CO563  PM10.11.CO564  PM10.11.CO565  PM10.11.CO566  PM10.11.CO567  PM10.11.CO568  PM10.11.CO569  PM10.11.CO570  PM10.11.CO571  PM10.11.CO572  PM10.11.CO573  PM10.11.CO574  PM10.11.CO575  PM10.11.CO576  PM10.11.CO577  PM10.11.CO578  PM10.11.CO579  PM10.11.CO580  PM10.11.CO581  PM10.11.CO582  PM10.11.CO583  PM10.11.CO584  PM10.11.CO585  PM10.11.CO586  PM10.11.CO587  PM10.11.CO588  PM10.11.CO589  PM10.11.CO590  PM10.11.CO591  PM10.11.CO592  PM10.11.CO593  PM10.11.CO594  PM10.11.CO595  PM10.11.CO596  PM10.11.CO597  PM10.11.CO598  PM10.11.CO599  PM10.11.CO600  PM10.11.CO601  PM10.11.CO602  PM10.11.CO603  PM10.11.CO604  PM10.11.CO605  PM10.11.CO606  PM10.11.CO607  PM10.11.CO608  PM10.11.CO609  PM10.11.CO610  PM10.11.CO611  PM10.11.CO612  PM10.11.CO613  PM10.11.CO614  PM10.11.CO615  PM10.11.CO616  PM10.11.CO617  PM10.11.CO618  PM10.11.CO619  PM10.11.CO620  PM10.11.CO621  PM10.11.CO622  PM10.11.CO623  PM10.11.CO624  PM10.11.CO625  PM10.11.CO626  PM10.11.CO627  PM10.11.CO628  PM10.11.CO629  PM10.11.CO630  PM10.11.CO631  PM10.11.CO632  PM10.11.CO633  PM10.11.CO634  PM10.11.CO635  PM10.11.CO636  PM10.11.CO637  PM10.11.CO638  PM10.11.CO639  PM10.11.CO640  PM10.11.CO641  PM10.11.CO642  PM10.11.CO643  PM10.11.CO644  PM10.11.CO645  PM10.11.CO646  PM10.11.CO647  PM10.11.CO648  PM10.11.CO649  PM10.11.CO650  PM10.11.CO651  PM10.11.CO652  PM10.11.CO653  PM10.11.CO654  PM10.11.CO655  PM10.11.CO656  PM10.11.CO657  PM10.11.CO658  PM10.11.CO659  PM10.11.CO660  PM10.11.CO661  PM10.11.CO662  PM10.11.CO663  PM10.11.CO664  PM10.11.CO665  PM10.11.CO666  PM10.11.CO667  PM10.11.CO668  PM10.11.CO669  PM10.11.CO670  PM10.11.CO671  PM10.11.CO672  PM10.11.CO673  PM10.11.CO674  PM10.11.CO675  PM10.11.CO676  PM10.11.CO677  PM10.11.CO678  PM10.11.CO679  PM10.11.CO680  PM10.11.CO681  PM10.11.CO682  PM10.11.CO683  PM10.11.CO684  PM10.11.CO685  PM10.11.CO686  PM10.11.CO687  PM10.11.CO688  PM10.11.CO689  PM10.11.CO690  PM10.11.CO691  PM10.11.CO692  PM10.11.CO693  PM10.11.CO694  PM10.11.CO695  PM10.11.CO696  PM10.11.CO697  PM10.11.CO698  PM10.11.CO699  PM10.11.CO700  PM10.11.CO701  PM10.11.CO702  PM10.11.CO703  PM10.11.CO704  PM10.11.CO705  PM10.11.CO706  PM10.11.CO707  PM10.11.CO708  PM10.11.CO709  PM10.11.CO710  PM10.11.CO711  PM10.11.CO712  PM10.11.CO713  PM10.11.CO714  PM10.11.CO715  PM10.11.CO716  PM10.11.CO717  PM10.11.CO718  PM10.11.CO719  PM10.11.CO720  PM10.11.CO721  PM10.11.CO722  PM10.11.CO723  PM10.11.CO724  PM10.11.CO725  PM10.11.CO726  PM10.11.CO727  PM10.11.CO728  PM10.11.CO729  PM10.11.CO730  PM10.11.CO731  PM10.11.CO732  PM10.11.CO733  PM10.11.CO734  PM10.11.CO735  PM10.11.CO736  PM10.11.CO737  PM10.11.CO738  PM10.11.CO739  PM10.11.CO740  PM10.11.CO741  PM10.11.CO742  PM10.11.CO743  PM10.11.CO744  PM10.11.CO745  PM10.11.CO746  PM10.11.CO747  PM10.11.CO748  PM10.11.CO749  PM10.11.CO750  PM10.11.CO751  PM10.11.CO752  PM10.11.CO753  PM10.11.CO754  PM10.11.CO755  PM10.11.CO756  PM10.11.CO757  PM10.11.CO758  PM10.11.CO759  PM10.11.CO760  PM10.11.CO761  PM10.11.CO762  PM10.11.CO763  PM10.11.CO764  PM10.11.CO765  PM10.11.CO766  PM10.11.CO767  PM10.11.CO768  PM10.11.CO769  PM10.11.CO770  PM10.11.CO771  PM10.11.CO772  PM10.11.CO773  PM10.11.CO774  PM10.11.CO775  PM10.11.CO776  PM10.11.CO777  PM10.11.CO778  PM10.11.CO779  PM10.11.CO780  PM10.11.CO781  PM10.11.CO782  PM10.11.CO783  PM10.11.CO784  PM10.11.CO785  PM10.11.CO786  PM10.11.CO787  PM10.11.CO788  PM10.11.CO789  PM10.11.CO790  PM10.11.CO791  PM10.11.CO792  PM10.11.CO793  PM10.11.CO794  PM10.11.CO795  PM10.11.CO796  PM10.11.CO797  PM10.11.CO798  PM10.11.CO799  PM10.11.CO800  PM10.11.CO801  PM10.11.CO802  PM10.11.CO803  PM10.11.CO804  PM10.11.CO805  PM10.11.CO806  PM10.11.CO807  PM10.11.CO808  PM10.11.CO809  PM10.11.CO810  PM10.11.CO811  PM10.11.CO812  PM10.11.CO813  PM10.11.CO814  PM10.11.CO815  PM10.11.CO816  PM10.11.CO817  PM10.11.CO818  PM10.11.CO819  PM10.11.CO820  PM10.11.CO821  PM10.11.CO822  PM10.11.CO823  PM10.11.CO824  PM10.11.CO825  PM10.11.CO826  PM10.11.CO827  PM10.11.CO828  PM10.11.CO829  PM10.11.CO830  PM10.11.CO831  PM10.11.CO832  PM10.11.CO833  PM10.11.CO834  PM10.11.CO835  PM10.11.CO836  PM10.11.CO837  PM10.11.CO838  PM10.11.CO839  PM10.11.CO840  PM10.11.CO841  PM10.11.CO842  PM10.11.CO843  PM10.11.CO844  PM10.11.CO845  PM10.11.CO846  PM10.11.CO847  PM10.11.CO848  PM10.11.CO849  PM10.11.CO850  PM10.11.CO851  PM10.11.CO852  PM10.11.CO853  PM10.11.CO854  PM10.11.CO855  PM10.11.CO856  PM10.11.CO857  PM10.11.CO858  PM10.11.CO859  PM10.11.CO860  PM10.11.CO861  PM10.11.CO862  PM10.11.CO863  PM10.11.CO864  PM10.11.CO865  PM10.11.CO866  PM10.11.CO867  PM10.11.CO868  PM10.11.CO869  PM10.11.CO870  PM10.11.CO871  PM10.11.CO872  PM10.11.CO873  PM10.11.CO874  PM10.11.CO875  PM10.11.CO876  PM10.11.CO877  PM10.11.CO878  PM10.11.CO879  PM10.11.CO880  PM10.11.CO881  PM10.11.CO882  PM10.11.CO883  PM10.11.CO884  PM10.11.CO885  PM10.11.CO886  PM10.11.CO887  PM10.11.CO888  PM10.11.CO889  PM10.11.CO890  PM10.11.CO891  PM10.11.CO892  PM10.11.CO893  PM10.11.CO894  PM10.11.CO895  PM10.11.CO896  PM10.11.CO897  PM10.11.CO898  PM10.11.CO899  PM10.11.CO900  PM10.11.CO901  PM10.11.CO902  PM10.11.CO903  PM10.11.CO904  PM10.11.CO905  PM10.11.CO906  PM10.11.CO907  PM10.11.CO908  PM10.11.CO909  PM10.11.CO910  PM10.11.CO911  PM10.11.CO912  PM10.11.CO913  PM10.11.CO914  PM10.11.CO915  PM10.11.CO916  PM10.11.CO917  PM10.11.CO918  PM10.11.CO919  PM10.11.CO920  PM10.11.CO921  PM10.11.CO922  PM10.11.CO923  PM10.11.CO924  PM10.11.CO925  PM10.11.CO926  PM10.11.CO927  PM10.11.CO928  PM10.11.CO929  PM10.11.CO930  PM10.11.CO931  PM10.11.CO932  PM10.11.CO933  PM10.11.CO934  PM10.11.CO935  PM10.11.CO936  PM10.11.CO937  PM10.11.CO938  PM10.11.CO939  PM10.11.CO940  PM10.11.CO941  PM10.11.CO942  PM10.11.CO943  PM10.11.CO944  PM10.11.CO945  PM10.11.CO946  PM10.11.CO947  PM10.11.CO948  PM10.11.CO949  PM10.11.CO950  PM10.11.CO951  PM10.11.CO952  PM10.11.CO953  PM10.11.CO954  PM10.11.CO955  PM10.11.CO956  PM10.11.CO957  PM10.11.CO958  PM10.11.CO959  PM10.11.CO960  PM10.11.CO961  PM10.11.CO962  PM10.11.CO963  PM10.11.CO964  PM10.11.CO965  PM10.11.CO966  PM10.11.CO967  PM10.11.CO968  PM10.11.CO969  PM10.11.CO970  PM10.11.CO971  PM10.11.CO972  PM10.11.CO973  PM10.11.CO974  PM10.11.CO975  PM10.11.CO976  PM10.11.CO977  PM10.11.CO978  PM10.11.CO979  PM10.11.CO980  PM10.11.CO981  PM10.11.CO982  PM10.11.CO983  PM10.11.CO984  PM10.11.CO985  PM10.11.CO986  PM10.11.CO987  PM10.11.CO988  PM10.11.CO989  PM10.11.CO990  PM10.11.CO991  PM10.11.CO992  PM10.11.CO993  PM10.11.CO994  PM10.11.CO995  PM10.11.CO996  PM10.11.CO997  PM10.11.CO998  PM10.11.CO999  PM10.11.CO1000  PM10.11.CO1001  PM10.11.CO1002  PM10.11.CO1003  PM10.11.CO1004  PM10.11.CO1005  PM10.11.CO1006  PM10.11.CO1007  PM10.11.CO1008  PM10.11.CO1009  PM10.11.CO1010  PM10.11.CO1011  PM10.11.CO1012  PM10.11.CO1013  PM10.11.CO1014  PM10.11.CO1015  PM10.11.CO1016  PM10.11.CO1017  PM10.11.CO1018  PM10.11.CO1019  PM10.11.CO1020  PM10.11.CO1021  PM10.11.CO1022  PM10.11.CO1023  PM10.11.CO1024  PM10.11.CO1025  PM10.11.CO1026  PM10.11.CO1027  PM10.11.CO1028  PM10.1
```

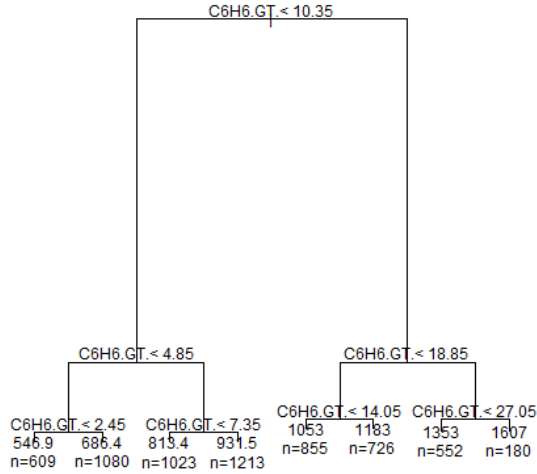
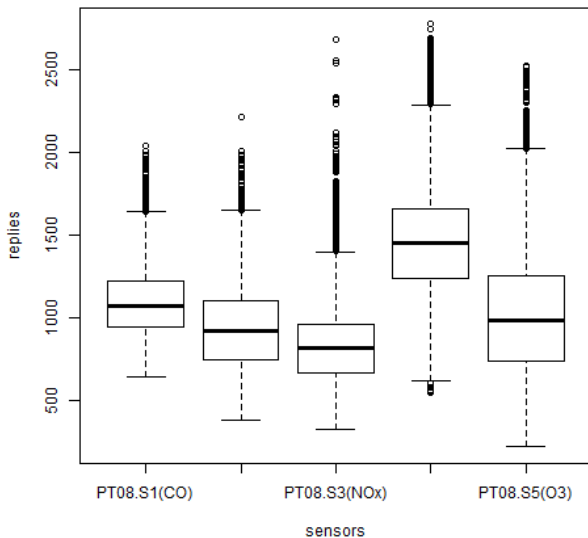


Figure 13: Árbol de decisión sobre la variable PT08.S2.NMHC.

#### 4. ANOVA

Se intentó finalmente realizar el test ANOVA sobre los datos para estudiar el factor sensor dado que todos los sensores miden algún tipo de óxido.

Primeramente se analizó la gráfica de medias y diagrama de cajas simultáneo como se muestra a continuación.



Las etiquetas corresponden a cada uno de los sensores; y el eje replies las concentraciones del compuesto correspondiente.

Se aprecia diferencia entre cada una de las medias de los sensores, así como la presencia de datos aislados

cuya cantidad es notable. Este fenómeno de la gran cuantía de datos aislados podemos afirmar que se debe a la sustitución de cada dato faltante por la media del resto. Estos datos remplazados se agrupan cerca de la media (porque tiene el mismo valor) y resulta que los datos reales quedan aislados. Se asegura entonces que el resultado del test va a arrojar diferencias entre las medias.

#### 4.1 Hipótesis y Modelo Estadístico

El modelo estadístico escogido fue el de Clasificación Simple. Tenemos claramente el factor sensor, el cual fue previamente seleccionado para saber si existe diferencia entre la concentración media de los compuestos.

¿Existe diferencia entre la concentración promedio medida por los sensores ?

La respuesta a esta pregunta es el resultado de contrastar las hipótesis:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

$$H_1 : \text{existen } 1 \leq i < j \leq 5 \text{ tales que } \mu_i \neq \mu_j$$

Donde  $\mu_i$  denota la concentración media medida por el sensor  $i$ . Como estudiamos en conferencias si denotamos  $y_{ij}$  como la concentración medida por el sensor  $i$  en la réplica  $j$ , la misma se puede escribir como:  $y_{ij} = \mu_i + e_{ij}$ , siendo  $e_{ij}$  el error experimental o la perturbación.

#### 4.2 Diferencias entre la concentración promedio medida por cada sensor

Realizamos la prueba de hipótesis planteada anteriormente para un nivel de significación no especificado. Nuevamente apoyándonos en el lenguaje R. Los resultados de la misma son los siguientes (figura 14).

```

> summary(sensors.anova)
      Df Sum Sq Mean Sq F value Pr(>F)
sensors  4 2.100e+09 524973580  5901 <2e-16 ***
Residuals 46780 4.162e+09  88963
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figure 14: Prueba de Hipótesis

Los resultados arrojan un  $p$ -value menor a  $2 \times 10^{-16}$ , el cual es prácticamente 0, pues el valor es cercano al menor número representable en la aritmética flotante de R. Con lo cual cualquier nivel de significación  $\alpha \in \{0.01, 0.1, 0.5\}$  es válido, con lo cual rechazamos la hipótesis nula y podemos afirmar que existe diferencia entre la concentración promedio medida por cada sensor.

La mayor concentración la tiene el 4to sensor. Dado el éxito de la prueba de hipótesis, partimos de la existencia de una diferencia, ahora bien, la evidencia visual nos muestra que la media del sensor 4 es superior, incluso si tenemos en cuenta los extremos del intervalo, estos también son superiores a los de los otros sensores.



### 4.3 Supuestos de normalidad y de igual varianza

La validez de los resultados obtenidos en cualquier análisis de varianza queda supeditada a que los supuestos del modelo se cumplan. Estos supuestos son:

1. Los  $e_{ij}$  siguen una distribución normal con media cero.
2. Los  $e_{ij}$  son independientes entre sí.
3. Los residuos de cada tratamiento tienen la misma varianza  $\sigma^2$ .

#### 4.3.1 PRUEBAS GRÁFICAS

Comenzando por las pruebas gráficas. Los resultados se muestran a continuación en las figuras 15, 16 y 17.

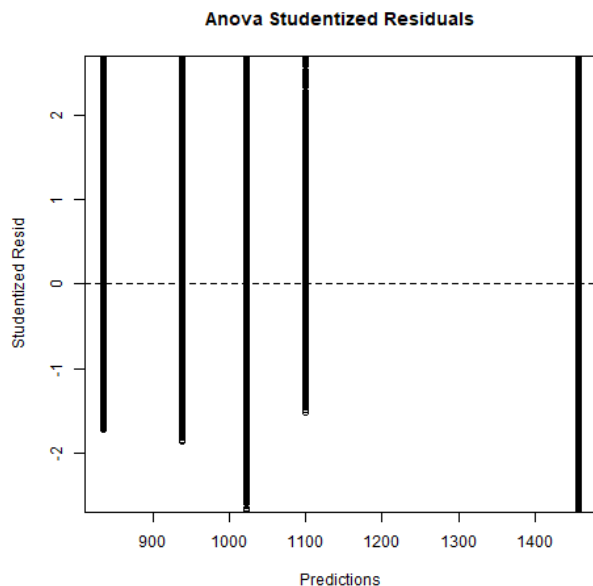


Figure 15: Residuos

Primeramente, en el gráfico estandarizado de residuos (figura 15), notamos los puntos muy dispersos, sin patrón aparente, por lo que podríamos asegurarnos el supuesto de varianza constante.

En el gráfico de predichos contra los residuos (figura 16), todos los puntos tienden a estar sobre la una misma recta, aunque no están completamente alineados, y esto se pudiera deber al la sustitución de los datos faltantes o a la cantidad considerable de datos. La cantidad de puntos aberrantes es muy pequeña con respecto al tamaño de los datos por lo que podemos tolerar y seguir afirmando que se cumple el supuesto de normalidad.

Finalmente, en la gráfica del histograma de residuos (figura 17), dicha gráfica se asemeja decentemente a la normal con media 0. Con lo cual afirmamos que se cumple el supuesto de normalidad.

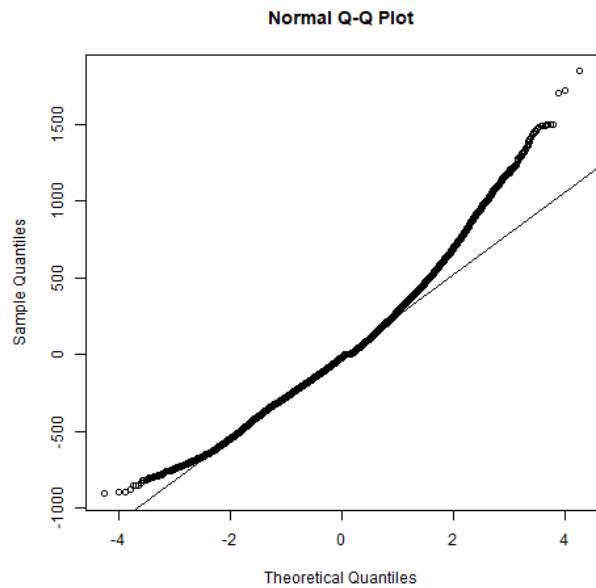


Figure 16: Predichos contra los residuos

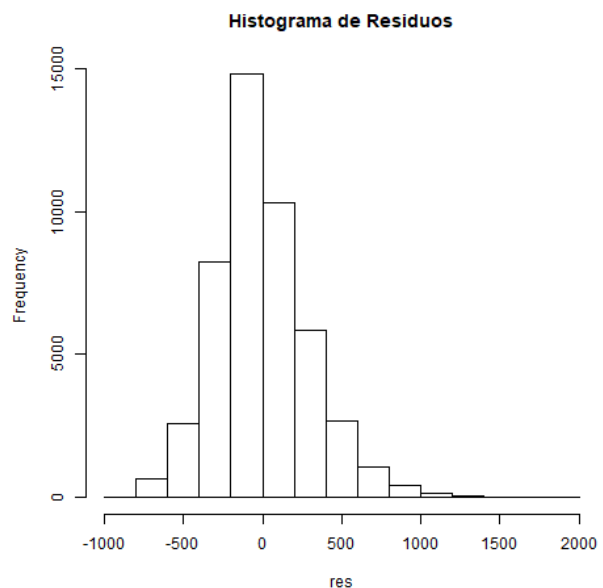


Figure 17: Histograma de residuos

#### 4.3.2 SHAPIRO-WILK, BARTLETT Y DURBIN-WATSON

El Test de Shapiro-Wilk no se pudo realizar por la gran cantidad de datos. Con lo cual, basándonos en las pruebas gráficas podemos asumir que el mismo se cumple y continuar con los dos test restantes.

```
Bartlett test of homogeneity of variances
data:  res and df$sensors
Bartlett's K-squared = 4486.3, df = 4, p-value < 2.2e-16
```

Figure 18: Test de igual varianza Bartlett

Como muestra la prueba de la figura 18 (Test de

Bartlett), el *p-value* muestra que es significativa, con lo que rechazamos  $H_0$ , y no se cumple el supuesto de varianzas constantes. La prueba deja de tener validez.

```
Durbin-watson test
data: sensors.anova
dw = 0.20914, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

Figure 19: Test Durbin-Watson

Finalmente, y de igual forma la prueba de la figura 19 (Test Durbin-Watson) muestra que el test es significativo, por tanto no se cumple el supuesto de independencia.

En conclusión, los dos últimos supuestos fallaron y por ende todos los resultados iniciales no son válidos. A pesar del estado de los datos, se realizó el análisis. Concluimos que estos datos no son lo idóneos para una prueba de ANOVA.

## 5. Conclusiones

Este equipo considera que la falta de valores presentes en la fuente impide que se pueda llegar a conclusiones específicas sobre los datos. Se recomienda realizar otra recopilación para un futuro estudio. Cualquier análisis con los datos actuales se considera como no confiable.

Todos los códigos, resultados e imágenes presentes en este documento y relacionados con los experimentos se encuentran adjuntos dentro del directorio del proyecto.