Fundamentos de Análisis de Datos
Trabajo Práctico 3

1.

Antes de realizar cualquier análisis es necesario validar nuestra muestra, en la Fig 1 podemos observar que en la variable choleste hay valores faltantes, entonces se debe definir el tratamiento de los mismos. Para este trabajo los valores faltantes se reemplazaron por la media de choleste según si son hombres o mujeres.

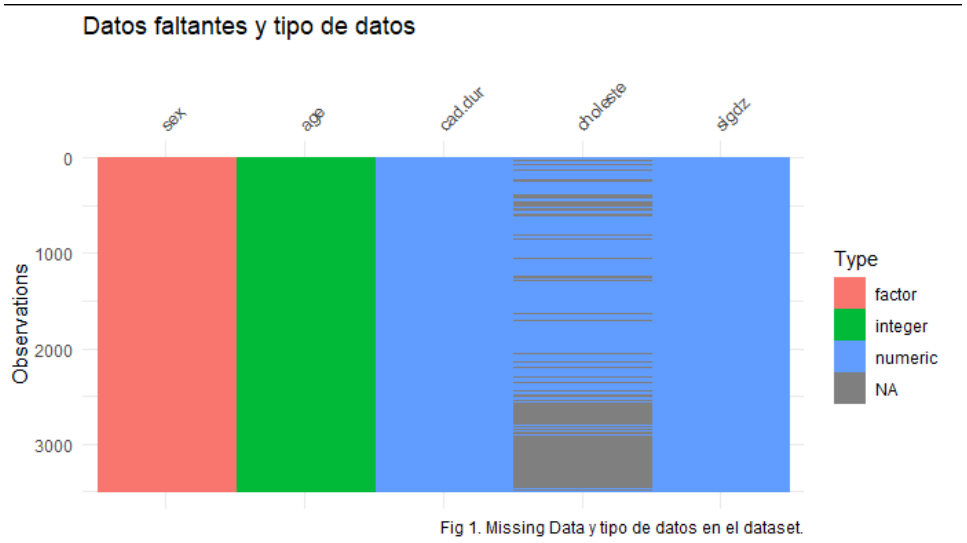

Fig 1. Missing Data y tipo de datos en el dataset.

La figura anterior nos muestra el tipo de dato y si los valores están faltantes o no, pero nada nos indica sobre la consistencia de los mismos, es decir que tengan valores lógicos. La Fig 2 muestra la distribución de las variables numéricas, en todas se observa gran cantidad de valores atípicos.

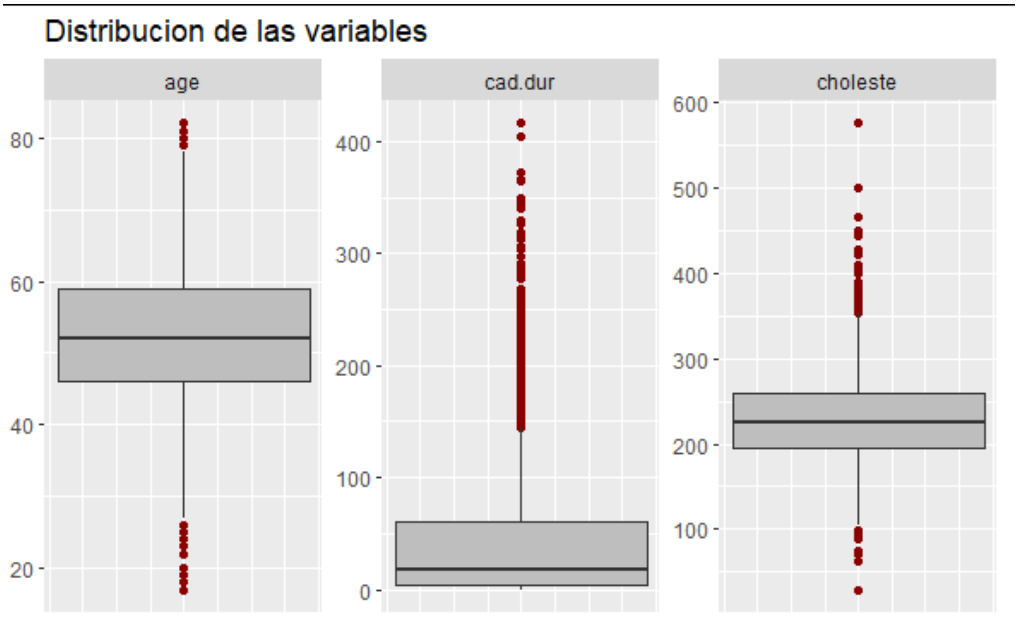

Fig 2. Boxplot de las variables

En este caso y dado el tipo de dataset, se decide dejar presente los valores atípicos ya que de eliminarlos se podría llegar a perder información sobre la relación entre la variable explicada y valores realmente fuera de la distribución de las variables explicativas.

Una vez validado el dataset es necesario conocer el modelo previo a su utilización. En este caso se utilizará un modelo de regresión logística, utilizado para modelar una variable binaria basado en una o más variables, llamadas predictoras.
Para este modelo se asume que

- Las variables predictoras no están correlacionadas entre sí
- Están significativamente relacionadas con la variable respuesta
- Las observaciones del modelo tampoco están correlacionados

El objetivo del modelo es estimar los verdaderos parámetros de la función densidad de probabilidad subyacente del modelo.

$$P(Y_i = 1) = F(\beta_0 + \beta_1 x_i + \varepsilon_i) \qquad (1)$$

Cuando la función F es una distribución logística como la siguiente, el modelo se denomina Regresión Logística

$$F(x) = \frac{e^x}{1 + e^x} \qquad (2)$$

Si reemplazamos (2) en (1) obtenemos:

$$P(Y_i = 1) = \frac{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n}}$$

Aplicando un poco de álgebra y logaritmo a ambos términos obtenemos la siguiente expresión, esta es una relación lineal llamada logit

$$ln\left(\frac{P(Y=1)}{1 - P(Y=1)}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

El término dentro del logaritmo se define como odds, y el de la derecha es igual a la expresión de una recta. Por ende, esta expresión nos permite relacionar la probabilidad $P(Y_i = 1)$ con un modelo lineal.

El modelo más simple que podemos crear con nuestro dataset es utilizar como variable explicada sigdz, y como variable explicativa choleste.
Se dividió el dataset inicial utilizando un 80% de los datos para entrenamiento, y un 20% para testeo.
Del modelo creado a partir de los datos de entrenamiento podemos observar la siguiente descripción en el comando summary:

```
Call:
glm(formula = sigdz ~ choleste, family = binomial, data = df_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8347  -1.4246   0.8620   0.8941   1.2847


Coefficients:
             Estimate   Std. Error   z value   Pr(>|z|)
(Intercept) -0.617445   0.247239    -2.497    0.0125 *
choleste     0.005852   0.001074     5.449    5.08e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 3543.8  on 2802  degrees of freedom
Residual deviance: 3512.4  on 2801  degrees of freedom
AIC: 3516.4

Number of Fisher Scoring iterations: 4
```

Del summary observamos que la variable explicativa choleste tiene un p valor muy bajo, por ende podemos decir que es estadísticamente significativa para el modelo. Además, dado estos resultados, y lo anteriormente expuesto sabemos que:

$$P(Y_i = 1) = \frac{e^{-0.562 + 0.00548 * choleste}}{1 + e^{-0.562 + 0.00548 * choleste}}$$

En la Fig 3 podemos observar la distribución logística en azul, los datos sigdz en función de la variable choleste y la predicción utilizando utilizando choleste = 199 en color rojo.
Si calculamos la probabilidad con la fórmula descrita anteriormente obtenemos:

$$P(Y_i = 1) = \frac{e^{-0.585 + 0.00556 * 199}}{1 + e^{-0.585 + 0.00556 * 199}} = 0.628$$

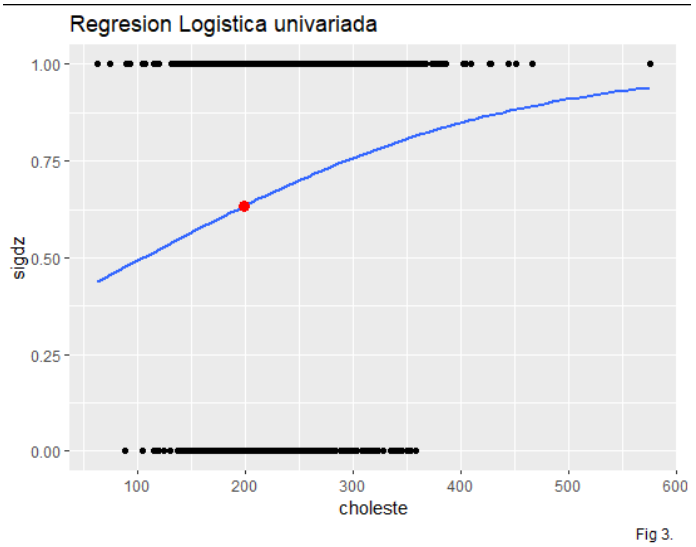

Fig 3.

Por último, si el modelo lo estamos utilizando para clasificar las observaciones y no solamente saber la probabilidad de P(Y=1), es interesante estudiar el desempeño del mismo utilizando una matriz de confusión de la Fig 4. Esta matriz se realizó utilizando el modelo ya creado y el dataset de testeo.

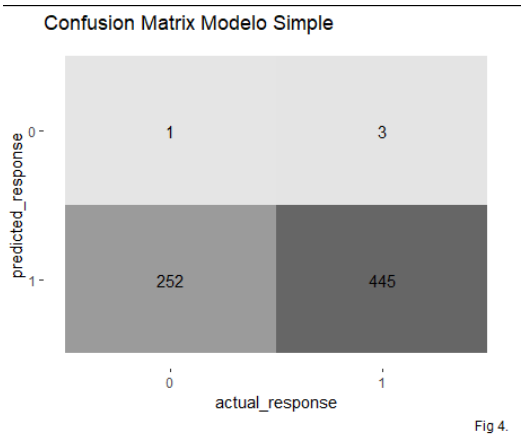

Fig 4.

Como métricas a utilizar podemos definir:
Accuracy : Las predicciones bien clasificadas sobre todas las predicciones.

$$\frac{TP+TN}{TP+TN+FN+TN}.$$

Precision: Mide los positivos verdaderos sobre todos los que dieron positivo

$$\frac{TP}{TP+FP}.$$

Recall : positivos verdaderos sobre todos los positivos

$$\frac{TP}{TP+FN}.$$

F1-score :

$$\frac{2 * Precision * Recall}{Precision + Recall}$$

En la Tabla 1 podemos observar los valores de las métricas calculadas:

Metricas Modelo Simple

	Metric	Estimate
1	accuracy	0.6362
2	precision	0.6385
3	recall	0.9933
4	f_meas	0.7773

Tabla 1

Si bien el recall del modelo es bueno, la precision y accuracy no lo son tanto. Esto se debe principalmente a que el modelo predice mayoritariamente 1 como valor de la variable dependiente.

Para tratar de mejorar las métricas de clasificación se crea un nuevo modelo teniendo en cuenta sólo las variables no categóricas del data set.
El comando summary utilizando este modelo da como resultado:

```
Call:
glm(formula = sigdz ~ choleste + age + cad.dur, family = binomial,
    data = df_train)

Deviance Residuals:
Min       1Q   Median       3Q      Max
-2.1902  -1.2375   0.7022   0.9026   1.6007

Coefficients:
             Estimate   Std. Error  z value   Pr(>|z|)
(Intercept) -3.3867850  0.3412582   -9.924   < 2e-16 ***
choleste     0.0057141  0.0010877    5.253   1.5e-07 ***
age          0.0542136  0.0046047   11.774   < 2e-16 ***
cad.dur      0.0003678  0.0008180    0.450     0.653
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 3543.8  on 2802  degrees of freedom
Residual deviance: 3342.7  on 2799  degrees of freedom
AIC: 3350.7

Number of Fisher Scoring iterations: 4
```

Como resultado de este summary podemos observar que dos de las tres variables son significativas (age, choleste) y la tercera cad.dur no lo es, sumado a que el score AIC se reduce frente al modelo anterior, pero no de una manera significativa.
La probabilidad de P(Y=1) en este modelo viene dado por la siguiente fórmula:

$$P(Y_i = 1) = \frac{e^{-0.562 + 0.00534 * choleste + 0.05566 * age + 0.000286 * cad.dur}}{1 + e^{-0.562 + 0.00534 * choleste + 0.05566 * age + 0.000286 * cad.dur}}$$

Nuevamente podemos utilizar las métricas para evaluar este modelo de clasificación, en la Fig 5 observamos la matriz de confusión del modelo, y en la Tabla 2 están las métricas anteriormente descritas. De esta manera podemos comparar el rendimiento del modelo contra el anterior.

Metricas Modelo No categoricas

	Metric	Estimate
1	accuracy	0.6633
2	precision	0.6677
3	recall	0.9420
4	f_meas	0.7815

Tabla 2

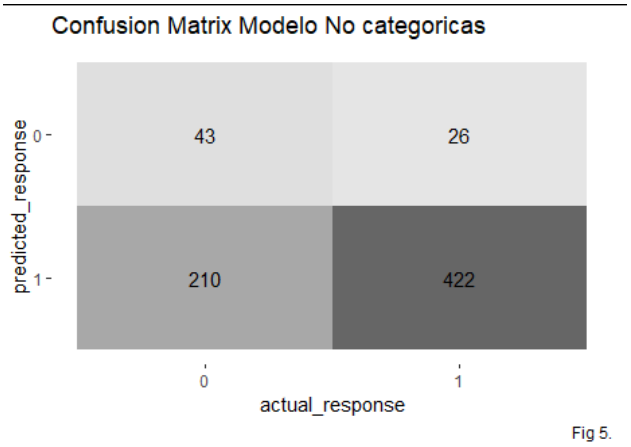

Fig 5.

En lo que respecta a las métricas se observa que este modelo tiene una ligera mejora de precisión y de accuracy respecto al modelo anterior, dado a que ahora no solamente predice valores 1 sino que se observa un incremento en la predicción de valores 0 para la variable explicada.
Y luego, la métrica F1 es ligeramente mayor a la del modelo simple.

Por último, se crea un modelo de regresión logística teniendo en cuenta la variable sex como factor sumado a las variables no categóricas anteriores.
Nuevamente el comando summary arroja la siguiente información:

```
Call:
glm(formula = sigdz ~ sex + choleste + age + cad.dur, family = binomial,
    data = df_train)

Deviance Residuals:
    Min      1Q  Median      3Q     Max
-2.5427  -0.8624   0.5123   0.7759   2.4771

Coefficients:
             Estimate   Std. Error  z value  Pr(>|z|)
(Intercept) -4.6864611  0.3857796  -12.148   < 2e-16 ***
sex1        -2.0777706  0.1031931  -20.135   < 2e-16 ***
choleste     0.0097307  0.0012334    7.889   3.04e-15 ***
age          0.0766283  0.0053125   14.424   < 2e-16 ***
cad.dur     -0.0002947  0.0008931   -0.330    0.741
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 3543.8 on 2802 degrees of freedom
Residual deviance: 2873.2 on 2798 degrees of freedom
AIC: 2883.2

Number of Fisher Scoring iterations: 4

Como se puede ver la única variable no significativa es cad.dur, la metrica AIC mejora considerablemente .
En tanto las métricas de clasificación obtenemos de la matriz de confusión de la Fig 6, la tabla con los valores mostrados en la Tabla 3

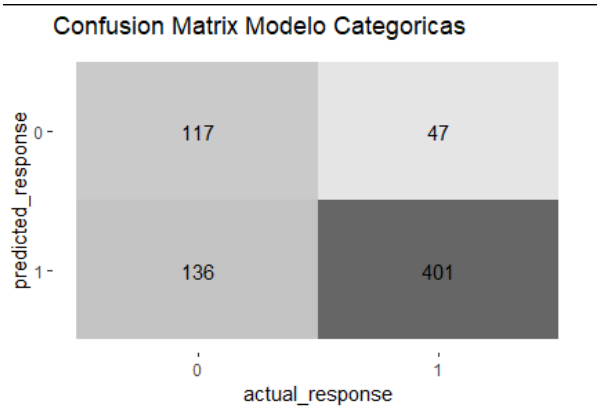

Fig 6.

Metricas Modelo Categoricas

	Metric	Estimate
1	accuracy	0.7389
2	precision	0.7467
3	recall	0.8951
4	f_meas	0.8142

Tabla 3

En este modelo observamos una mejora considerable en las métricas accuary y precision frente a los dos modelos anteriores, en desmedro de recall. Pero esta disminución de recall, se ve superado por la mejora en las primeras métricas, dando así una mejora sustancial en la métrica F1.

Teniendo en cuenta el valor score AIC, y las métricas de clasificación podemos concluir que el mejor modelo es el último, el que tiene la variable sex como factor.

2.

El test ANOVA es un tipo de test estadístico usado para determinar si hay una diferencia estadísticamente significativa entre la media de dos o más grupos.

Para realizar este test se realizan los siguientes supuestos:

- Todo valor observado de la variable dependiente puede expresarse mediante la siguiente función:

$$y_{ij} = \beta_0 + m_i + \varepsilon_{ij} \quad ; con\, i = 1, \ldots, k\,; \quad y\, j = 1,\ldots, n$$

- Dentro del grupo las observaciones deben ser independientes; y entre grupos, estos deben ser independientes uno del otro.

- Aproximación a la normalidad, las variables respuesta deben ser aproximadamente normales.

- Los grupos deberían tener aproximadamente la misma varianza.

En este test se definen dos hipótesis, en la hipótesis nula todas las medias de los grupos son iguales; y en la hipótesis alternativa existe al menos una que es diferente a las demás.
Si el p valor observado en el test es menor al nivel de significancia que se eligió entonces se rechaza la hipótesis nula, y se valida la alternativa.

El estadístico estudiado en ANOVA es conocido como F_{ratio} y se calcula de la siguiente manera:

$$F_{ratio} = \frac{varianza\ de\ las\ medias\ entre\ los\ grupos}{promedio\ de\ la\ varianza\ dentro\ de\ los\ grupos}$$

Si se cumple la hipótesis nula, el estadístico F_{ratio} resulta 1 ya que la intervarianza es igual a la intravarianza. Entonces cuanto mayor sea la diferencia de la media de los grupos mayor será el valor de F_{ratio}.

En el ejercicio planteado se tienen tres muestras con frecuencias de asistencia a clases teóricas en tres localidades distintas, y se desea estudiar si existen diferencias estadísticamente significativas entre la cantidad de estudiantes que asisten.

Con el fin de conocer y chequear la consistencia de los datos y su distribución. Analizaremos los datos gráficamente en la Fig 1, y descriptivamente como se muestra a continuación:

```
$CAS
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6.00   10.00   11.00   11.60   14.00   20.00

$POD
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    4.00    6.00    5.45    7.00   11.00

$SL
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0     4.0     7.5     6.9    10.0    16.0
```

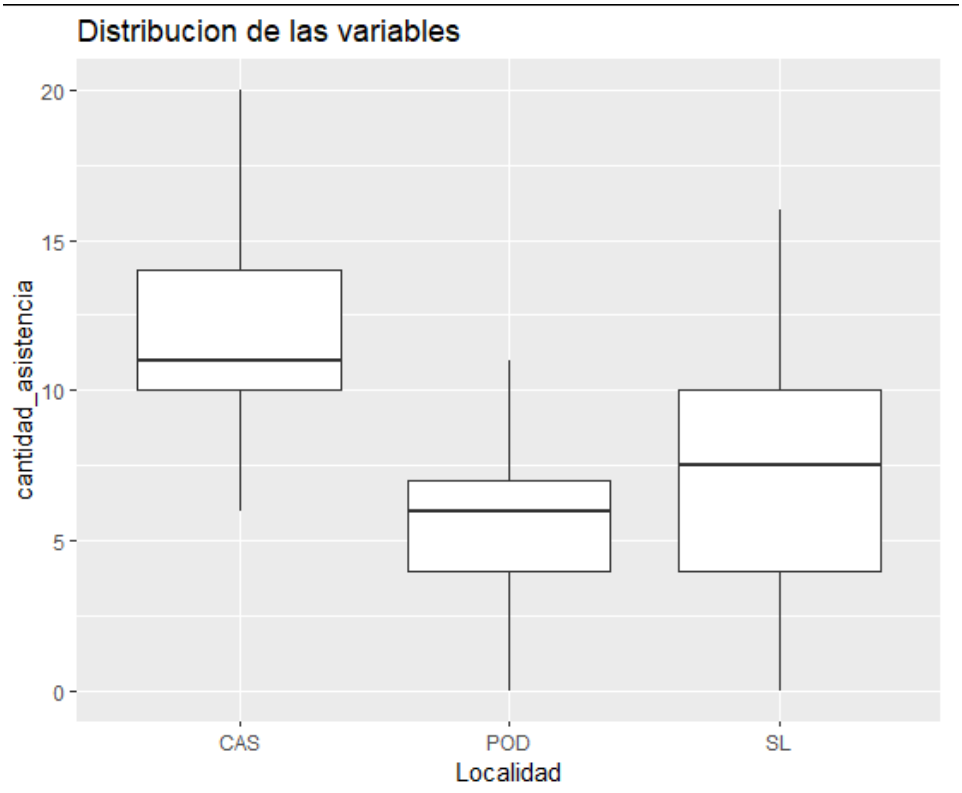

Fig 1. Boxplot variables

Tanto en la descripción de las distribuciones como en el gráfico se puede observar una diferencia entre la media y la mediana del primer grupo con respecto a la de los demás, y también una menor variabilidad del segundo grupo. Pero para poder confirmar si existe una diferencia significativa es necesario realizar un test ANOVA.

Para realizar este test se deben plantear dos hipótesis:

$$H_0: \textit{La media de todos los grupos son iguales}$$
$$H_1: \textit{Existe al menos una media que es distinta a las demás.}$$

Al realizar este test se obtienen los siguientes valores:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
localidad	2	564.2	282.10	20.26	2.78e-08	***
Residuals	117	1629.1	13.92			

Dado que el p valor es realmente bajo, existe evidencia suficiente para rechazar la hipótesis nula y confirmar la alternativa.
Una vez realizado el modelo, es necesario validar el mismo. En la Fig 2 vemos tres gráficos para validar el modelo, el primero y el tercero utilizado para validar la igualdad de la varianza intergrupo utilizando los residuos y los residuos estandarizados respectivamente, y un QQ plot para validar la normalidad de los residuos.
De los gráficos de los residuos se observa que un grupo posee menor variabilidad, sumado a que los residuos tienen una distribución que no se ajusta a la normal según el QQ plot, esto es importante a

la hora de sacar conclusiones ya que ambos son supuestos del modelo. La salida del shapiro test se muestra a continuación, y corrobora la no normalidad de los residuos:

```
Shapiro-Wilk normality test

data:  anova_test$residuals
W = 0.98473, p-value = 0.1945
```

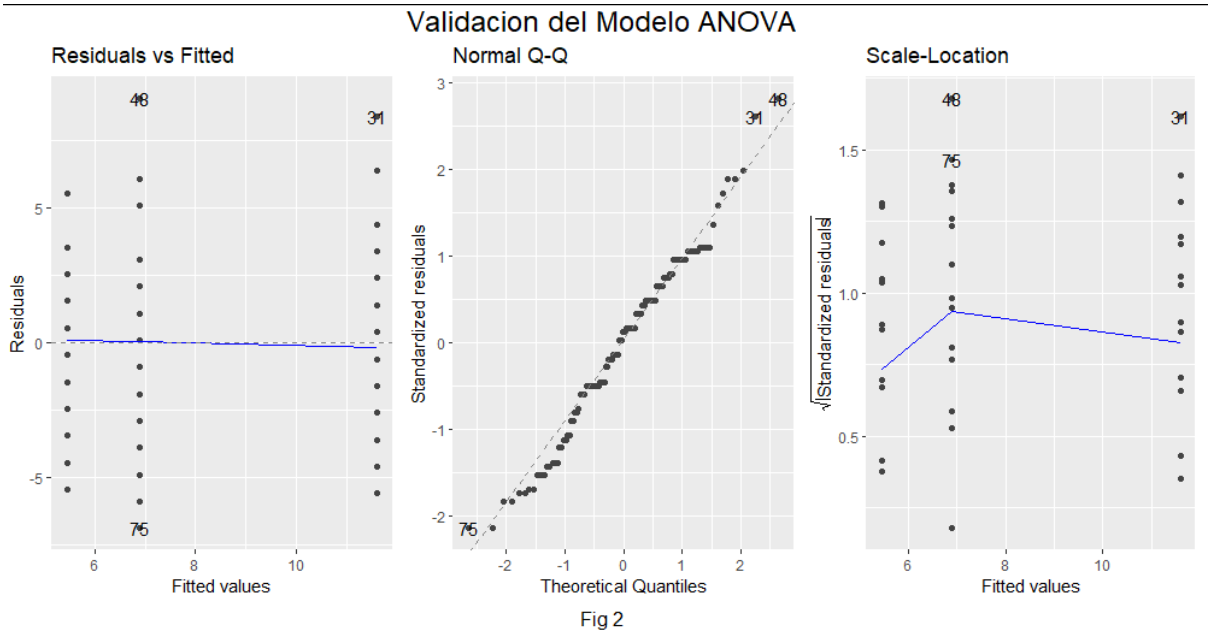

Fig 2

La limitante que tiene este test de ANOVA es la de identificar cual es la variable que tiene media distinta, simplemente nos indica que hay una diferencia. Para identificar esta diferencia se debe realizar un test HSD de Tukey. La Fig 3 muestra la diferencia entre las medias de a pares, si el intervalo de confianza cruza por el 0 no se podrá afirmar que ambas distribuciones tienen medias distintas.

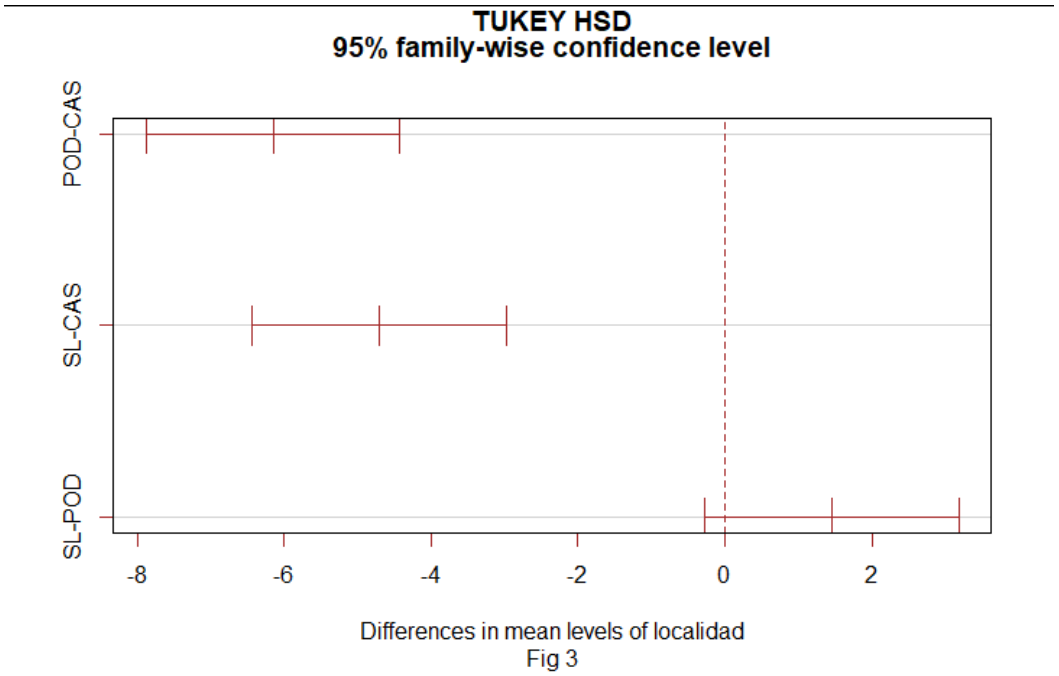

Fig 3

Además, veremos a continuación la salida del test de Tukey para validar con estadísticos y p valores las conclusiones que podemos sacar del gráfico.

```
Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = frecuencia_asistencia ~ localidad)

$localidad
             diff       lwr        upr     p adj
POD-CAS -5.15 -7.1307536 -3.169246 0.0000000
SL-CAS  -3.70 -5.6807536 -1.719246 0.0000620
SL-POD   1.45 -0.5307536  3.430754 0.1955537
```

Tanto del gráfico como de la salida del test podemos observar que la localidad Podestá y Caseros tienen distinta media con un p valor extremadamente bajo, y también la media entre las localidades Santos Lugares y Caseros son distintas con un p valor extremadamente bajo.
No se puede decir lo mismo de la media entre la localidad de Santos Lugares y Podestá que podrían ser iguales.

3.

Una serie de tiempo es una secuencia de observaciones obtenidas en determinados momentos de tiempo, está ordenada cronológicamente y espaciada de manera uniforme. El objetivo de analizar una serie de tiempo es realizar pronósticos. Para ello es necesario construir un modelo que describa de una manera sencilla la evolución de la serie a través del tiempo.
La serie del archivo de datos de R “nottem” contiene las temperaturas promedio por mes en la ciudad de Nottingham, en la Fig. 1 se puede observar el gráfico de esta serie de tiempo.
Esta serie comienza en enero de 1920 con una temperatura promedio de 40.6 ºF, y termina en diciembre de 1939 con una temperatura promedio de 37.8 ºF.

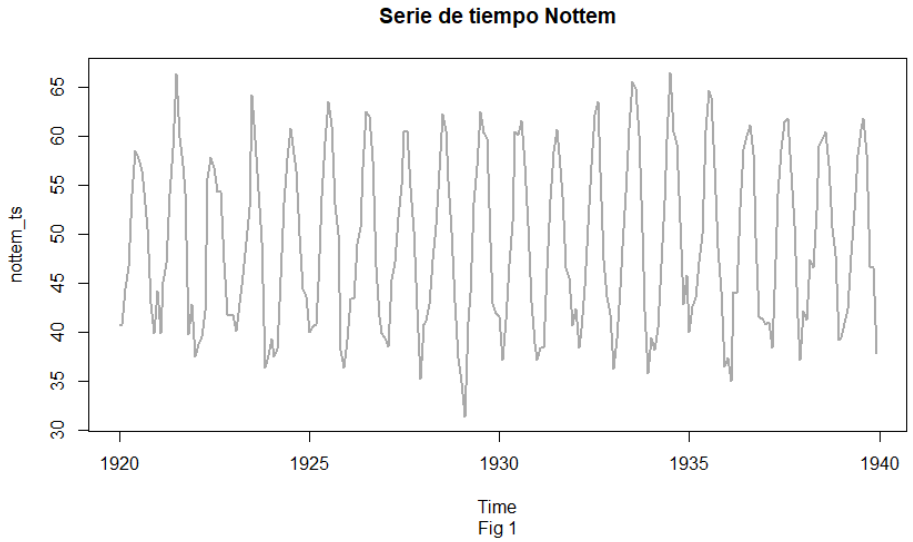

Fig 1

Para saber si una serie de tiempo es estacionaria se puede realizar el test ADF con un nivel de significancia de 0.05 cuyo resultado podemos observar a continuación:

```
        Augmented Dickey-Fuller Test

data:  nottem_ts
Dickey-Fuller = -12.998, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary
```

Como el p valor es mejor al nivel de significancia se define la serie como estacionaria. Es importante que la serie sea estacionaria ya que los modelos para realizar predicciones serán mucho más precisos.

Una serie de tiempo puede descomponerse en una componente tendencia, una estacional y una aleatoria. En la Fig 2, se observa la componente tendencia y en la Fig. 3 se observa un gráfico boxplot
que comprueba la estacionalidad de la serie. Si no hubiera estacionalidad los boxplot deberían tener una distribución similar.

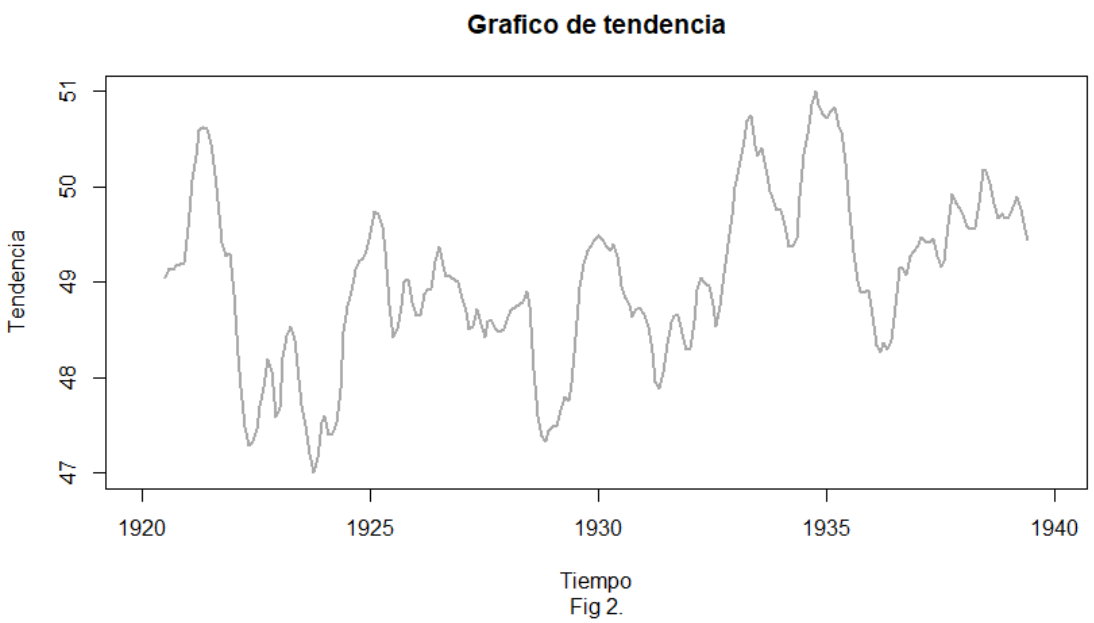

Fig 2.

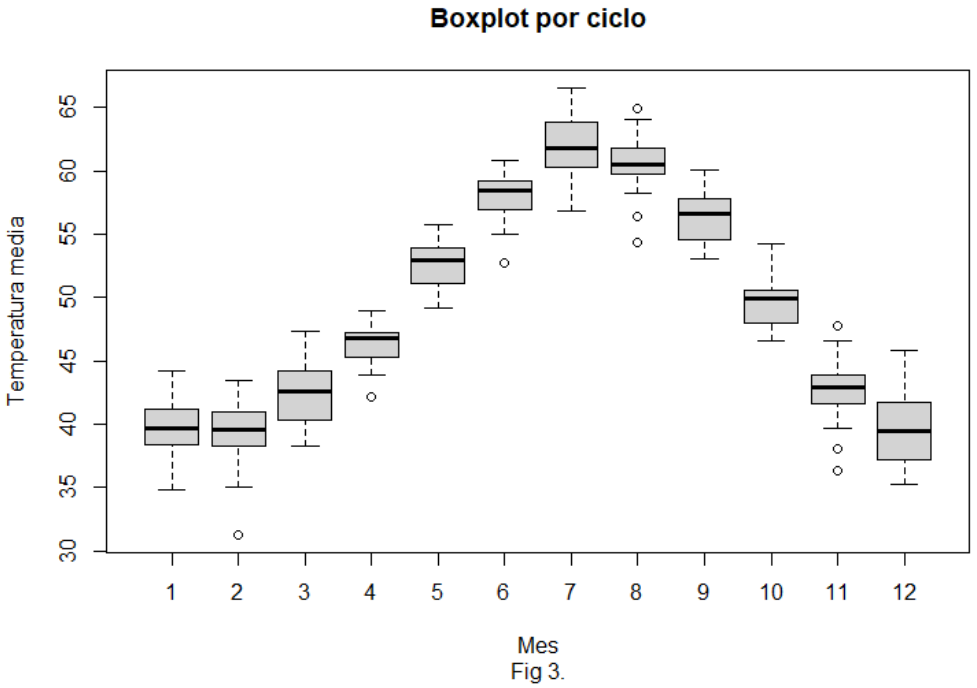

Fig 3.

Es posible construir un modelo ARIMA que posibilite el forecast de la serie de tiempo. Un modelo ARIMA se compone de una parte autoregresiva, una integración y una parte de media móvil que determina el número de términos de los tiempos anteriores incluidos en el modelo. Dada las tres componentes se utilizan tres parámetros: 'p', 'd' , 'q'.
El parámetro 'p' representa la cantidad de observaciones anteriores con la que está relacionada la observación actual, el parámetro 'd' representa el grado de diferenciación en la componente integrada y por último, el parámetro 'q' determina el número de términos de tiempos anteriores incluidos en el modelo.
Para el análisis de la serie de tiempo nottem, se creó un modelo cuyos parámetros se muestran a continuación, y así posibilitar la realización de un forecast de la serie:

Series: AP
ARIMA(1,0,2)(1,1,2)[12] with drift

Coefficients:

	ar1	ma1	ma2	sar1	sma1	sma2	drift
	0.1526	0.1225	0.1100	-0.5181	-0.5447	-0.2066	0.0041
s.e.	0.3759	0.3716	0.1229	0.1742	0.1882	0.1744	0.0041

sigma^2 = 5.236: log likelihood = -490.45
AIC=996.91 AICc=997.6 BIC=1023.91

Para crear el modelo anterior se dividió la serie en dos, utilizando todos los años menos el último para el dataset de entrenamiento, y el último año como dataset de testeo y así comparar los valores forecast con los reales.
En la Fig 4 se muestra el forecast creado con el modelo anteriormente descrito, se resalta en azul la media de los valores pronosticados. Y en la Tabla 1 podemos ver los valores actuales del dataset de testeo y los valores forecast, y su correspondiente error.

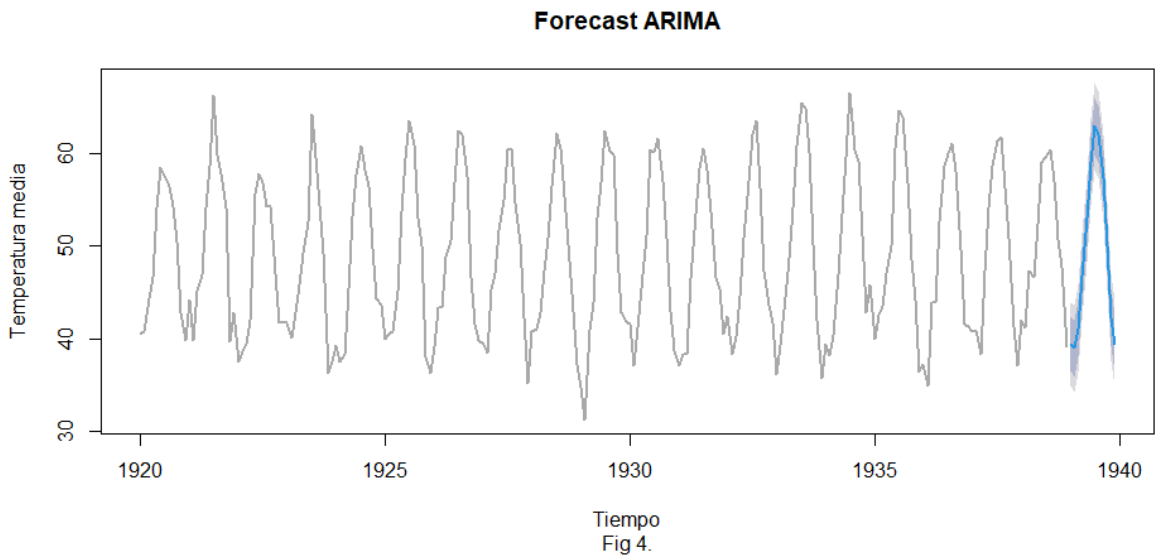

Fig 4.

Actual vs Forecasted

	Mes	Actual	Forecasted	Error
1	Jan	39.4	39.5	-0.1
2	Feb	40.9	39.0	1.9
3	Mar	42.4	41.1	1.3
4	Apr	47.8	46.8	1.0
5	May	52.4	53.0	-0.6
6	Jun	58.0	59.1	-1.1
7	Jul	60.7	62.9	-2.2
8	Aug	61.8	62.1	-0.3
9	Sep	58.2	57.3	0.9
10	Oct	46.7	50.0	-3.3
11	Nov	46.6	42.4	4.2
12	Dec	37.8	39.5	-1.7

Tabla 1