Ejercicios de Análisis Exploratorio de Datos con Tidyverse

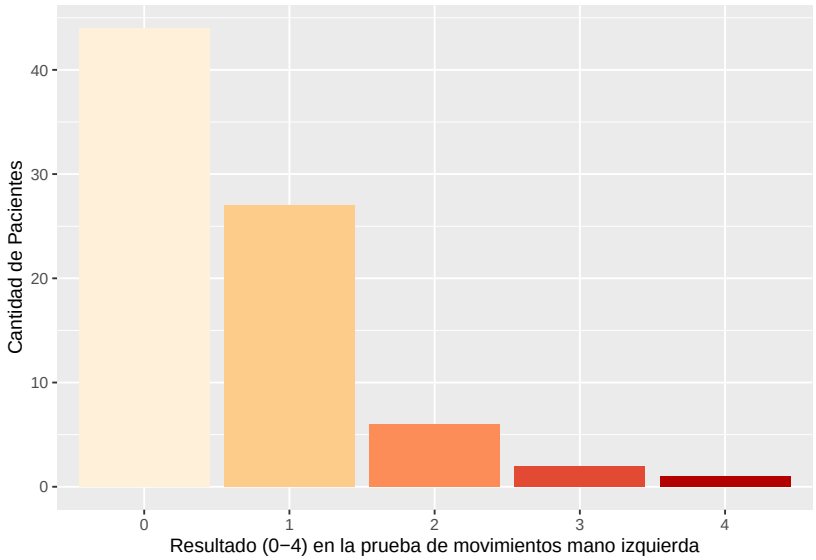
Pau Vives, Arnold Cruz, Samuel de Paúl

22/10/2021

Visualización e interpretación de variables:

1. Una de las variables cualitativas de vuestro conjunto de datos (1 punto)

La variable que interpretaremos será *X24.LUE*, ésta variable indica el resultado (de 0 a 4) obtenido en la prueba de movimientos en la mano izquierda, realizada para evaluar a los pacientes en la escala UPDRSIII. Los valores varían entre 0 (Ausente) y 4 (De gran amplitud, interfiere la alimentación). Utilizaremos un gráfico de barras para visualizar la distribución de la variable.



Contamos también cuantos valores hay en cada barra, para tener el máximo de información posible:

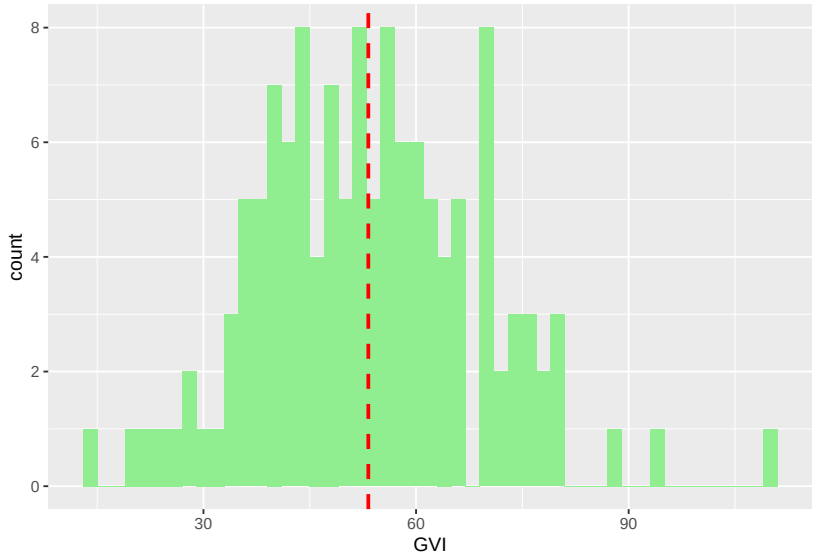
```
## # A tibble: 5 x 2
##   `X24.LUE[1:80]`     n
##   <chr>           <int>
## 1 0                  44
## 2 1                  27
## 3 2                   6
## 4 3                   2
## 5 4                   1
```

Por tanto, como vemos, la mayoría de los pacientes no presentan movimientos o temblores en la mano izquierda, o presentan movimientos ligeros; mientras que una pequeña fracción de los pacientes tiene problemas serios en lo que a ésta prueba respecta.

2. Una de las variables cuantitativas de vuestro conjunto de datos. (1 punto)

Analizaremos la variable cuantitativa llamada *GVI*. Ésta variable corresponde a una medición realizada a

los pacientes mientras éstos leían un texto fonéticamente equilibrado. Se mide la separación media entre los intervalos de voz.



A continuación daremos los valores más representativos de la dispersión de los datos:

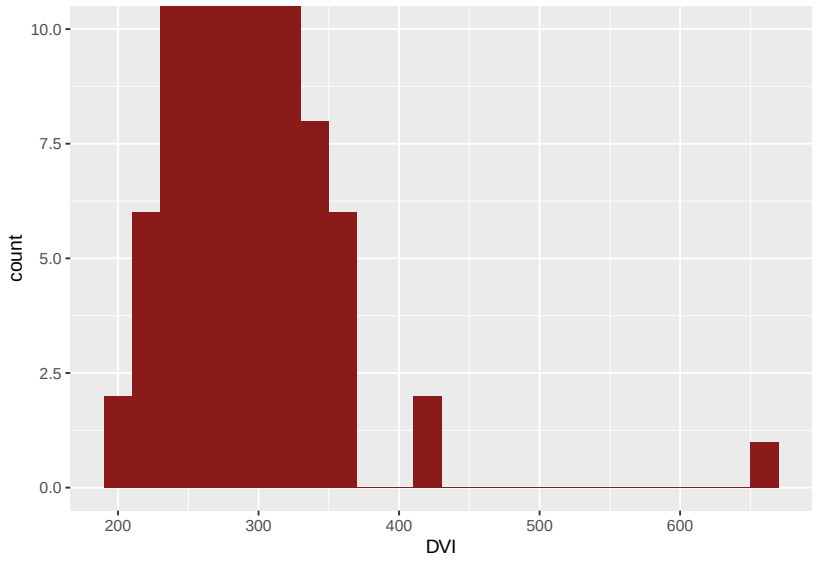
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.72   42.71   52.67   53.26   62.95  109.50
```

```
## [1] 15.60616
```

Vemos por tanto que los datos, pese a tener un valor mínimo y máximo bastante lejanos, están agrupados en torno a la media, que es de 53.26, con una desviación típica de 15.6. Vemos valores bastante alejados de los demás, tanto por lo alto como por lo bajo.

3. Una de las variables que presente un patrón inusual y/o valores atípicos. (2 puntos)

Buscamos una variable que tenga valores inusuales representándola mediante un histograma, acercando la imagen a los valores más pequeños del eje vertical.



Observamos que en la variable *DVI* tenemos algún outlier. Al igual que en el caso anterior, la variable se

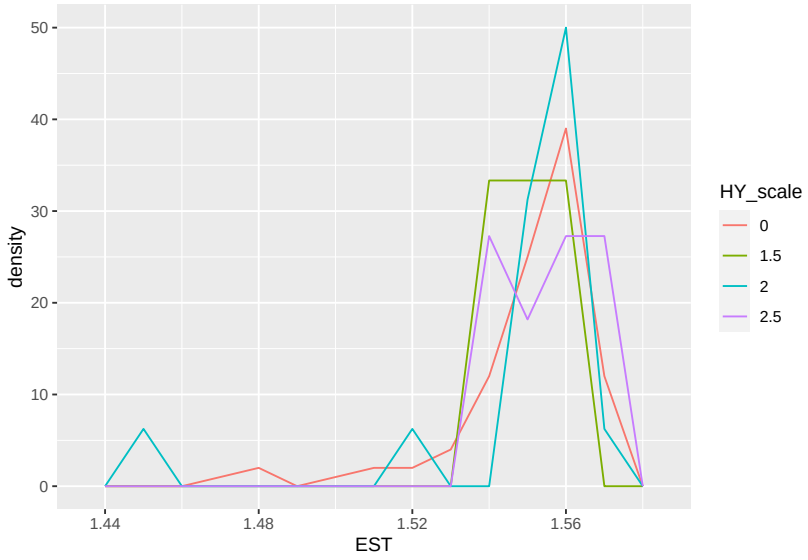
corresponde a una medición realizada a los pacientes mientras éstos leían un texto fonéticamente equilibrado. Se mide la duración media de los intervalos de voz. La gran mayoría de los valores se encuentran entre 200 y 400, mientras que, en el histograma observamos que hay valores que están por encima de 650. Veamos ahora cuáles son esos valores.

```
## # A tibble: 1 x 6
##   code    age gender HY_scale UPDRSIII   DVI
##   <chr> <int> <chr>  <chr>    <chr>    <int>
## 1 PD04     75 M      2        24         663
```

Vemos que se trata del paciente PD04, de 75 años, que no tiene valores execsivamente altos ni de *UPDRSIII* (escala generalizada en la que se mide la gravedad del Parkinson) ni de \textit{HY_scale} (un sistema de uso común para describir cómo progresan los síntomas de la enfermedad de Parkinson), por lo que no encontramos un motivo claro para éste valor tan aislado.

4. El análisis conjunto de una variable categórica y una cuantitativa. (2 puntos)

Para este apartado tomaremos la variable categórica \textit{HY_scale}, cuyo valor representa en qué fase del avance de los síntomas del Parkinson se encuentra cada paciente, y la variable cualitativa *EST*, que mide la heterogeneidad en el habla en términos de la ocurrencia de intérvalos de sonoridad, insonoridad, pausas y respiraciraciones. Los pacientes de la muestra que no tienen Parkinson están marcados con "-". Sustituíremos este símbolo por el 0, que será equivalente y más cómodo de manejar.



Observamos que los valores correspondientes a los pacientes con puntuación más alta en la escala H&Y (2.5) se situan en un rango bastante alto de *EST*, como era de esperar. A su vez, vemos que la línea roja, correspondiente al nivel 0 en la escala (son pacientes que no tienen Parkinson) concentra más valores que el resto de niveles en la zona de baja *EST*. Vemos también que los niveles 1.5 y 2 tienen curvas de densidad similares a las demás, aunque concentran la mayoría de sus valores en un rango de *EST* ligeramente menor al del nivel 2.5.