PAD - Guía 1 - Problema 1

Grupo 1

2025-04-14

a)

- https://www.rdocumentation.org/packages/MASS/versions/7.3-65/topics/UScereal

Los datos corresponden a mediciones de la información nutricional de distintos cereales presentes en supermercados de Estados Unidos (Tabla 1). Cada observación representa el contenido nutricional de un producto en específico.

Tabla 1. Descripción de las variables presentes en el archivo UScereal.

Variable	Descripción	Tipo de dato	Tipo de variable
`mfr`	(manufacturer) Fabricante del cereal (A = American Home Foods, G = General Mills, etc.)	Factor 6 lvls	Categórica
`calories`	Calorías por porción	double	Cuantitativa continua
`protein`	Gramos de proteína por porción	double	Cuantitativa continua
`fat`	Gramos de grasa por porción	double	Cuantitativa continua
`sodium`	Miligramos de sodio por porción	double	Cuantitativa continua
`fibre`	Gramos de fibra dietética por porción	double	Cuantitativa continua
`carbo`	Gramos de carbohidratos complejos por porción	double	Cuantitativa continua
`sugars`	Gramos de azúcares por porción	double	Cuantitativa continua
`shelf`	Número de estante en el supermercado (1, 2 o 3, de abajo hacia arriba)	int	Categórica ordinal
`potassium`	Gramos de potasio por porción	double	Cuantitativa continua
`vitamins`	Vitaminas y minerales (none, enriched, 100)	Factor 3 lvls	Categórica

¿Cómo considera que se realizó la recolección de datos?

Una duda que nos surgió es si estos datos se tomaron como una muestra, o si constituyen el total de la población de cereales disponibles en el mercado, con lo que constituiría un censo. Podrían haber elegido supermercados al azar en distintas zonas y tomar los datos de los productos de cereales a la venta. Entendemos en este caso que el objetivo era analizar los cereales a la venta en Estados Unidos, dado que no se aclara ningún estado o localidad en particular. Tal vez el objetivo era comprobar si la información nutricional cumplía las normas de la FDA, o conocer más acerca de los nutrientes que aportan los cereales a la venta en supermercados.

Dado que la distribución de las variables categóricas `mfr` y `vitamins` es tan desbalanceada consideramos que no son factores que se hayan tenido en cuenta al momento de tomar la muestra. Por otro lado, la variable `shelf` parece un factor de muestreo más plausible, no solo por su distribución más balanceada, sino también porque es razonable hipotetizar que la posición en el estante se vea influida por atributos nutricionales como el contenido de azúcar.

En cualquier caso, tenemos por certeza que los cereales se examinaron en un supermercado, no se eligió más de un ejemplar por producto, y todos los datos extraídos, excepto `shelf` dependen de la marca o el fabricante.

Otros datos:

- ASA: This stands for the American Statistical Association. It's a professional organization for statisticians. The ASA promotes the practice and profession of statistics. They are very involved in the development and distribution of statistical knowledge.
- F&DA: This refers to the Food and Drug Administration (FDA). It's a United States federal agency responsible for protecting and promoting public health through the control and supervision of food safety, tobacco products, pharmaceuticals, medical devices, and other products. In the context of the "UScereal" dataset, the "mandatory F&DA food label" refers to the nutritional information that food manufacturers are required by the FDA to provide on their product packaging.

b)

Para realizar un ANOVA, las variables respuesta lógicas serían las variables continuas del conjunto de datos: `calories`, `protein`, `fat`, `sodium`, `fibre`, `carbo`, `potassium` y `sugars`.

Si pensáramos en **factores** como variables categóricas, podríamos considerar inicialmente `mfr`, `shelf` y `vitamins`.

Sin embargo, al examinar la distribución de `vitamins` y `mfr`, observamos un fuerte desbalance:

- **vitamins:** La gran mayoría de las observaciones (57) pertenecen a la categoría "enriched", con muy pocas en "none" (3) y "100%" (5).

```
  100% enriched     none
     5       57        3
```

- **mfr:** Aunque algo menos extremo que `vitamins`, también presenta categorías con pocas observaciones (N=3, Q=5, R=5) en comparación con otras (G=22, K=21).

```
 G  K  N  P  Q  R
22 21  3  9  5  5
```

Este desbalance en `mfr` y `vitamins` dificulta la aplicación del ANOVA por varias razones:

1. Es complicado verificar adecuadamente los supuestos de normalidad y homocedasticidad en los niveles del factor con tan pocas observaciones.

2. Los resultados pueden ser muy sensibles a valores atípicos en esos grupos pequeños.
3. Se obtiene poca información fiable sobre la distribución de la variable respuesta en esos niveles.

Por estas razones, no sería recomendable utilizar `mfr` o `vitamins` como factores en un ANOVA con estos datos. Surge la opción, en todo caso, de recurrir a transformaciones de los niveles presentes en alguna de estas categorías, tomando el riesgo de perder cierta información en el proceso.

Por su parte, la variable **`shelf`** presenta una distribución más equilibrada entre sus niveles. Además, si el análisis exploratorio previo (ej. mediante boxplots) sugiere posibles diferencias en las variables continuas entre los distintos estantes, sería interesante y metodológicamente más robusto realizar un ANOVA utilizando `shelf` como factor para investigar formalmente esas diferencias.

La elección de factores o efectos en el modelo debería tener en consideración alguna interpretación práctica de los potenciales resultados. En este sentido, resultaría interesante estudiar si hubo presentación diferencial de cereales de una determinada calidad en góndolas más accesibles a la vista (para este caso, góndola 3), dándoles una posición preferencial de frente al consumidor.

c) a)

Robert O. Kuehl en su libro "Diseño de Experimentos" define tratamiento y factor de la siguiente manera:

- **Tratamiento**: "Los tratamientos son el conjunto de circunstancias creados para el experimento, en respuesta a la hipótesis de investigación y son el centro de la misma. Entre los ejemplos de tratamientos se encuentran dietas de animales, producción de variedades de cultivos, temperaturas, tipos de suelo y cantidades de nutrientes." (p. 4)
- **Factor**: "Un factor es un grupo específico de tratamientos: temperatura, humedad y tipos de suelo se consideran un factor cada uno. Las diversas categorías de un factor se denominan niveles del factor." (p. 7)

Por su parte no propone una definición clara de efecto fijo si no que lo define en contraste con los efectos aleatorios:

	Factor fijo	Factor aleatorio
Niveles del factor	Todos los niveles del factor que son de interés se incluyen en el estudio.	En el experimento se incluye una muestra aleatoria de los posibles niveles del factor.
Inferencia	La inferencia estadística no tiene alcance más allá de los niveles estudiados.	La inferencia estadística se orienta hacia todos los posibles niveles del factores.

De acuerdo a lo expuesto por este autor podemos entender que cada nivel de `mfr` constituiría un tratamiento diferente y el conjunto de ellos sería el factor fabricante. Luego, considerar a `mfr` como factor fijo implica que, por un lado proporciona datos necesarios para encontrar respuesta a las preguntas que generaron el estudio, y por otro que todos los niveles del factor que son de interés están incluidos. Esto implicaría que las conclusiones de análisis no se extienden más allá de los niveles considerados.

La definición de factor fijo depende en parte de los objetivos del estudio, los cuales desconocemos. Si bien no sabemos si se incluyeron todos los niveles de interés, con los datos disponibles podemos plantear un experimento (o suponerlo) y considerar el factor en estudio como fijo para responder una pregunta específica. Esto último presenta el problema de que no sabemos cómo se recolectaron los datos, ni las decisiones que se tomaron, por lo tanto nuestras conclusiones siempre se verán limitadas por ello.

Por último, cabe aclarar que el término de factor o efecto fijo no tiene una única definición, aquí partimos nuestra reflexión acerca de la propuesta de Robert O. Kuehl, sin embargo, si hubiésemos partido de otra definición nuestra respuesta podría haber sido muy diferente. Al respecto de las distintas definiciones es interesante el siguiente link: https://statmodeling.stat.columbia.edu/2005/01/25/why_i_dont_use/

c) b)

Indicar cuál sería el objetivo al que responde el ANOVA en este caso.

Seleccionamos la variable `potassium` como variable respuesta. El objetivo principal de aplicar un ANOVA en este caso, utilizando la variable continua 'potassium' (contenido de potasio) y la variable categórica 'mfr' (fabricante), es determinar si existen diferencias estadísticamente significativas en el contenido medio de potasio entre los cereales producidos por los distintos fabricantes presentes en el dataset.

En esencia, se busca responder a la pregunta: ¿Varía el nivel promedio de potasio en los cereales dependiendo de qué empresa los fabrica? Al observar el dataset, notamos variabilidad considerable en los niveles de potasio (desde valores tan bajos como 15 hasta valores superiores a 900 g), lo que sugiere que podrían existir diferencias significativas entre fabricantes. El ANOVA nos ayudaría a determinar si estas diferencias observadas son estadísticamente significativas.

Un análisis como este podría tener por detrás un objetivo práctico, por ejemplo, guiar a consumidores que buscan cereales con mayor o menor contenido de potasio, o conocer si ciertos fabricantes se especializan o se distinguen en cuanto a la cantidad de este mineral en sus productos.

Por último, como todas las variables continuas que tenemos se extrajeron de la información nutricional del producto, en cualquier caso estaríamos analizando qué fabricante se diferencia de los otros en la cantidad de ese nutriente, o si no existe ninguna diferencia.

¿Considera adecuado realizar este análisis dada la manera en que se recolectan los datos? Justifique su respuesta.

Si nuestra suposición acerca de la recolección de datos es correcta, es decir, si se seleccionaron al azar uno o más supermercados y de allí se tomaron cajas de cereales de los distintos niveles de las góndolas, tomando una por cada producto específico, entonces `mfr` no fue una variable controlada en el estudio. Por lo tanto, los fabricantes presentes en la muestra pueden considerarse una muestra aleatoria de la población de fabricantes de cereales disponibles. En este escenario, el objetivo del análisis no sería comparar los fabricantes específicos presentes en la muestra, sino inferir sobre la variabilidad entre la población de fabricantes. Esto llevaría a considerar a `mfr` como un factor aleatorio (de acuerdo a la definición propuesta más arriba).

Nuestro desconocimiento sobre las condiciones del muestreo realizado siembra dudas en cuanto a si las distintas marcas fueron homogéneamente incluidas en el estudio, o si hubo preferencia por algunas empresas específicas.

Escriba el modelo lineal, indicado lo que significa cada uno de sus términos.

El modelo a ajustar sería:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Siendo

- Y_{ij} es la variable respuesta `potassium`

- μ es la media general
- α_i es el efecto del nivel i del factor `mfr`
- ε_{ij} es el error aleatorio.
- $i \in [1, 6]$, por los 6 niveles del factor `mrf`
- El valor de j depende de las observaciones en cada nivel, es decir $j_1, \ldots, j_k$ dado que las cantidad de observaciones es distinta para cada empresa.

Para los términos asociados a variables aleatorias indique cuáles son los supuestos subyacentes.

Una de las variables aleatorias del modelo son los errores o residuos (la diferencia entre cada observación y la media de su grupo), estos deben cumplir con algunos supuestos para que se garantice que se pueda justificar el uso de la distribución F para realizar inferencias sobre las diferencias entre grupos. Estos supuestos son:

- **Independencia:** Los errores (y por lo tanto, las observaciones) deben ser independientes entre sí. Esto significa que el valor de una observación no debe influir ni estar relacionado con el valor de otra observación. Esto se garantiza generalmente mediante un muestreo aleatorio adecuado.
- **Normalidad:** Los errores (o residuos) deben seguir una distribución normal con media cero dentro de cada grupo. El ANOVA es relativamente robusto a violaciones moderadas de este supuesto, especialmente con tamaños de muestra grandes, gracias al Teorema Central del Límite.
- **Homocedasticidad (Homogeneidad de Varianzas):** La varianza de los errores debe ser constante en todos los grupos. Es decir, las poblaciones de las que se extraen las muestras para cada grupo deben tener la misma varianza ($\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$).

El otro término asociado a una variable aleatoria es la variable respuesta. El supuesto de normalidad de los errores implica que la variable respuesta debe ser normal *condicionada a cada grupo* (dentro de cada grupo), pero no implica que la variable respuesta total (agrupando todos los datos sin distinguir por grupo) deba ser normal.

Plantee las hipótesis a probar, en términos de los parámetros del modelo.

Para el efecto α queremos probar los siguiente:

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = 0$$
$$H_1 : \text{ al menos una } \alpha_i \neq 0$$

Complete la siguiente Tabla ANOVA

Fuentes de variación	Grados de libertad	Suma de cuadrados	Cuadrado medio
Fabricante	6-1=5	315991.7	63198.34
Error	65-6=59	1764263	29902.76
Total	65-1=64	2080254	

¿cómo se obtiene el valor del Estadístico de la prueba y cuál es su distribución?

Consideramos que SC1 = 315991.7 y que SC2 = 1764263. El valor estadístico de la prueba se obtiene a partir de obtener los respectivos cuadrados medios, i.e. dividir cada suma de cuadrados por sus respectivos

grados de libertad (presentados en la misma fila en b-vi). Posteriormente, estos cuadrados medios se dividen entre sí.

$$CM_m = SC_m/k_m$$

Siendo

- CM_m el cuadrado medio del m-ésimo término del modelo
- SC_m la suma de cuadrados de dicho término
- k_m los grados de libertad de dicho término

$$F_{obs} = CM_\alpha/CM_{error}$$

Para este caso el estadístico F_{obs} presenta un valor de 2.1135, en una distribución F de parámetros 5 y 59.

Análisis descriptivo, análisis del modelo y conclusiones

```
data = UScereal
str(data)
```

```
## 'data.frame':    65 obs. of  11 variables:
##  $ mfr      : Factor w/ 6 levels "G","K","N","P",..: 3 2 2 1 2 1 6 4 5 1 ...
##  $ calories : num  212 212 100 147 110 ...
##  $ protein  : num  12.12 12.12 8 2.67 2 ...
##  $ fat      : num  3.03 3.03 0 2.67 0 ...
##  $ sodium   : num  394 788 280 240 125 ...
##  $ fibre    : num  30.3 27.3 28 2 1 ...
##  $ carbo    : num  15.2 21.2 16 14 11 ...
##  $ sugars   : num  18.2 15.2 0 13.3 14 ...
##  $ shelf    : int  3 3 3 1 2 3 1 3 2 1 ...
##  $ potassium: num  848.5 969.7 660 93.3 30 ...
##  $ vitamins : Factor w/ 3 levels "100%","enriched",..: 2 2 2 2 2 2 2 2 2 2 ...
```

```
data$shelf = as.factor(data$shelf)
```

```
skim(data)
```

Table 4: Data summary

Name	data
Number of rows	65
Number of columns	11
Column type frequency:	
factor	3
numeric	8
Group variables	None

Variable type: factor

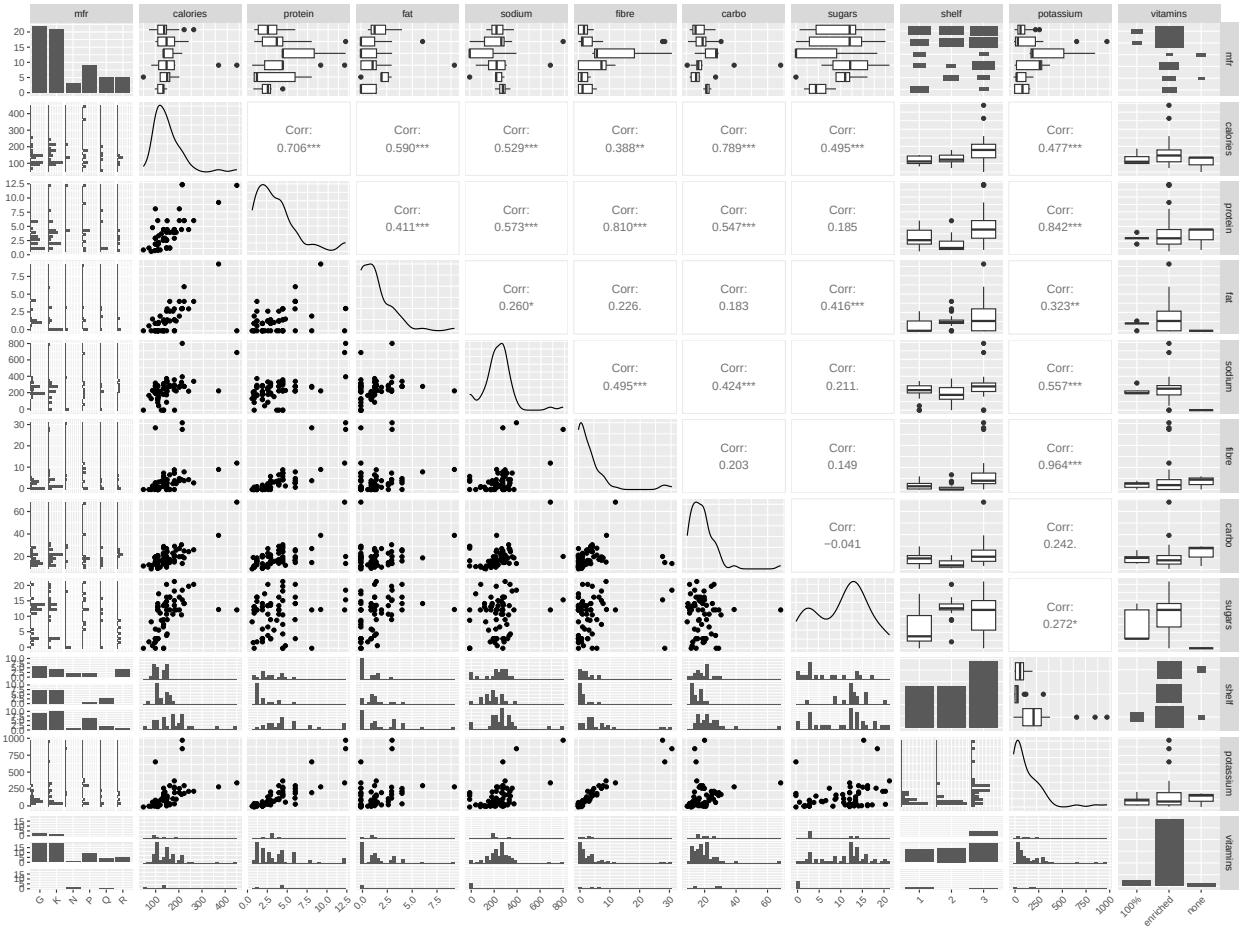
skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
mfr	0	1	FALSE	6	G: 22, K: 21, P: 9, Q: 5
shelf	0	1	FALSE	3	3: 29, 1: 18, 2: 18
vitamins	0	1	FALSE	3	enr: 57, 100: 5, non: 3

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
calories	0	1	149.41	62.41	50.00	110	134.33	179.10	440.00	
protein	0	1	3.68	2.64	0.75	2	3.00	4.48	12.12	
fat	0	1	1.42	1.65	0.00	0	1.00	2.00	9.09	
sodium	0	1	237.84	130.63	0.00	180	232.00	290.00	787.88	
fibre	0	1	3.87	6.13	0.00	0	2.00	4.48	30.30	
carbo	0	1	19.97	8.47	10.53	15	18.67	22.39	68.00	
sugars	0	1	10.05	5.84	0.00	4	12.00	14.00	20.90	
potassium	0	1	159.12	180.29	15.00	45	96.59	220.00	969.70	

Tenemos 65 datos y 11 variables, de las cuales 3 son categóricas, dos de ellas (`mfr` y `vitamins`) están muy desbalanceadas. Con respecto a nuestra variable de interés, `potassium` vemos que tiene un rango muy grande, con un mínimo en 15 y un máximo en 969.70. Además, la desviación estándar es 180.29, la cual es más alta que la media 159.12 lo cual es un indicio de mucha variabilidad.

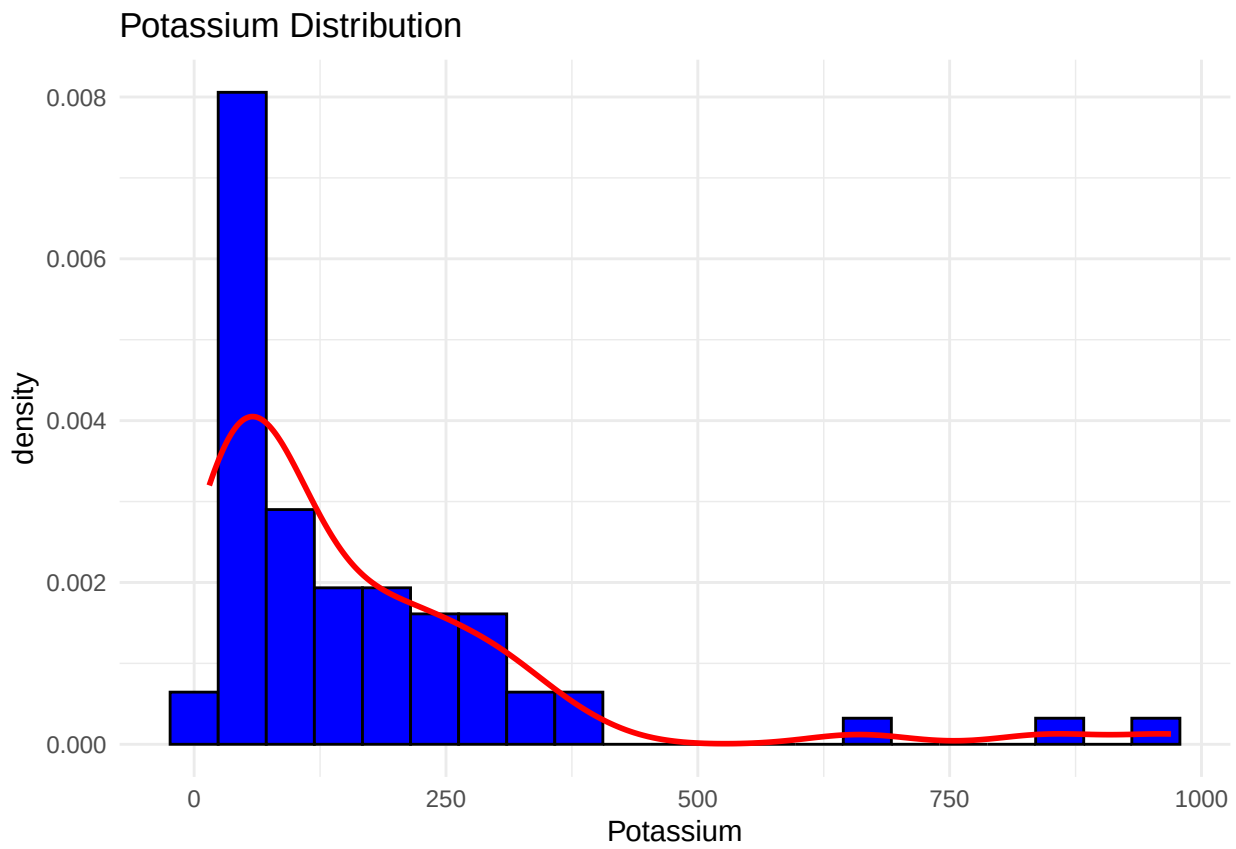
```
p = ggpairs(data, progress = FALSE) + theme(axis.text.x = element_text(angle = 45, hjust = 1))
print(p)
```




```
ggsave("pairs_plot.png", p, width = 15, height = 15, dpi = 300)
```

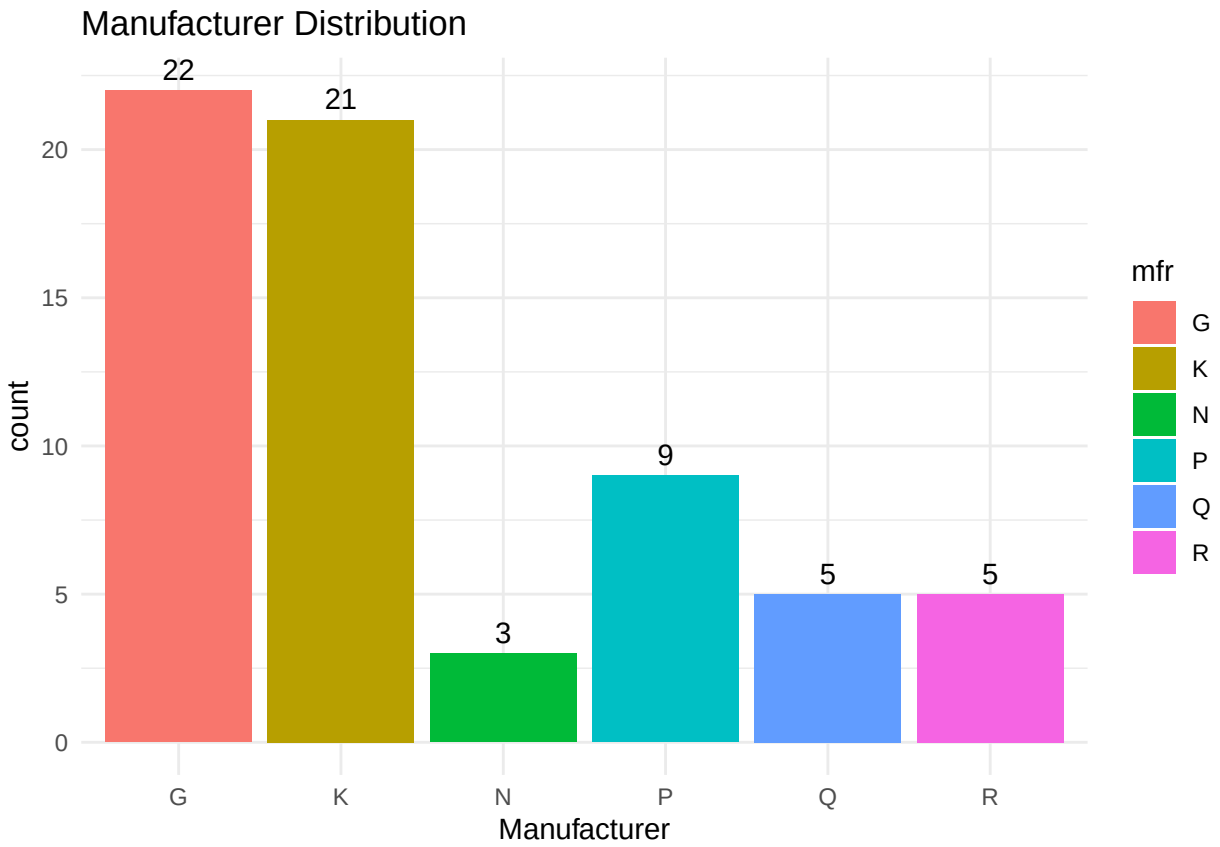
Observando las distintas distribuciones podemos decir que las variables numéricas son en su mayoría asimétricas, mostrando en sus distribuciones un par de valores atípicos. En particular potassium presenta una asimetría derecha muy pronunciada.

```
p1 <- ggplot(data, aes(x = potassium)) +
  geom_histogram(aes(y = after_stat(density)), fill="blue", binwidth = (max(data$potassium, na.rm=TRUE)
  geom_density(color="red", linewidth=1) +
  ggtitle("Potassium Distribution") +
  xlab("Potassium") +
  theme_minimal()
p1
```



En este gráfico es mucho más notorio el sesgo a la derecha de la variable `potassium`. También observamos algunos posibles valores atípicos muy altos, los cuales explican la variabilidad observada.

```
# distribución mfr histogram
ggplot(data, aes(x = mfr, fill = mfr)) +
  geom_bar() +
  geom_text(stat = "count", aes(label = after_stat(count)), vjust = -0.5) +
  ggtitle("Manufacturer Distribution") +
  xlab("Manufacturer") +
  theme_minimal()
```



Los niveles del factor mfr están desbalanceados, en la gráfica se ve que algunos fabricantes tienen muchas observaciones (G, K), mientras que otros tienen muy pocas (N, P, Q, R). El desbalance puede afectar la potencia estadística y la forma en que se calculan los efectos (sobre todo si se usan métodos de sumas de cuadrados tipo I, II o III). Se desconoce si este desbalance corresponde a la amplitud de productos presentados por algunas de las empresas o a un muestreo sesgado por parte de quienes confeccionaron la base de datos.

```
data %>%
  group_by(mfr) %>%
  skim(potassium)
```

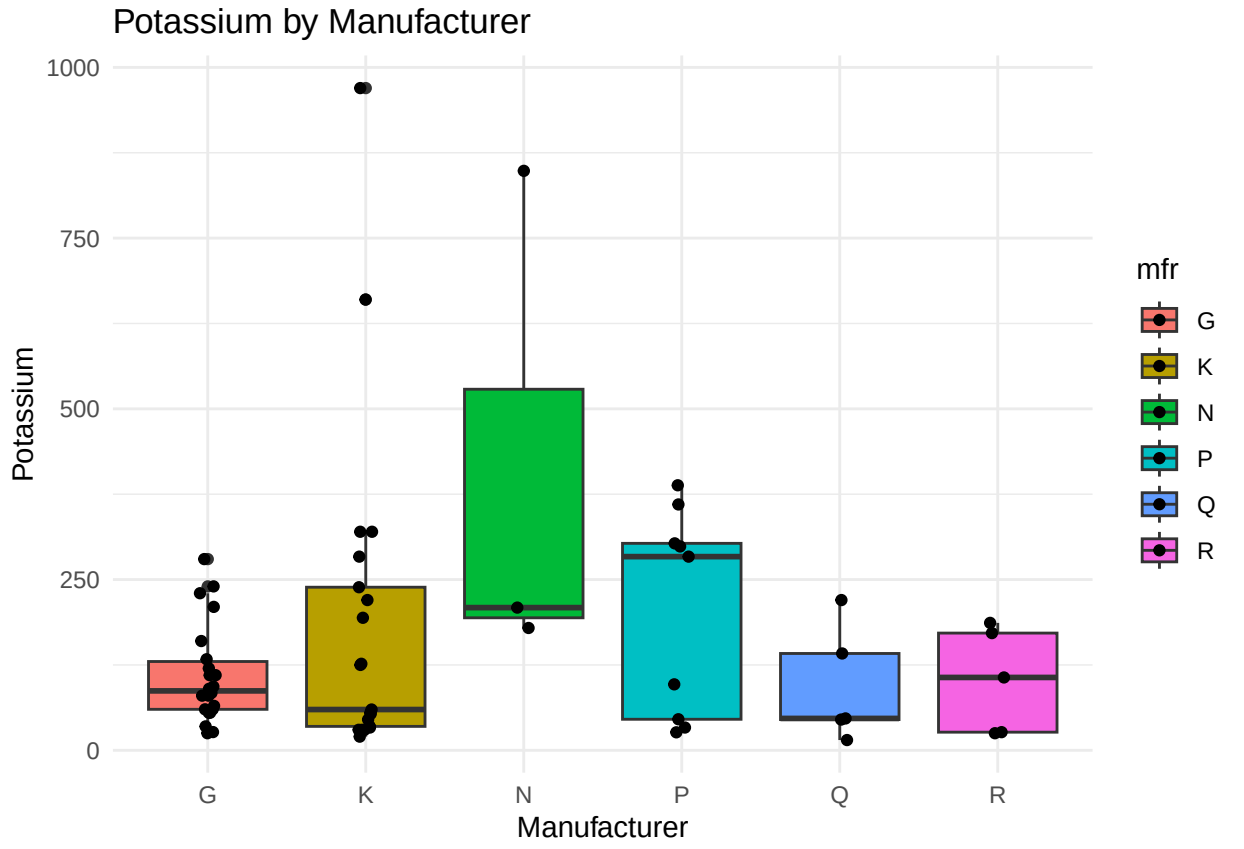
Table 7: Data summary

Name	Piped data
Number of rows	65
Number of columns	11
Column type frequency:	
numeric	1
Group variables	mfr

Variable type: numeric

skim_variable	mfr	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
potassium	G	0	1	109.20	72.16	25.00	60.00	87.00	130.00	280.00	
potassium	K	0	1	184.96	238.28	20.00	35.00	59.70	238.81	969.70	
potassium	N	0	1	412.18	378.14	179.10	194.03	208.96	528.72	848.48	
potassium	P	0	1	203.87	150.27	26.32	45.45	283.58	303.03	388.06	
potassium	Q	0	1	93.69	85.21	15.00	45.00	46.67	141.79	220.00	
potassium	R	0	1	103.28	76.87	25.00	26.55	106.67	171.64	186.57	

```
# Distribución de la variable `potassium` por fabricante
ggplot(data, aes(x = mfr, y = potassium, fill = mfr)) +
  geom_boxplot() +
  geom_point(position = position_jitterdodge(jitter.width = 0.4)) +
  ggtitle("Potassium by Manufacturer") +
  xlab("Manufacturer") +
  ylab("Potassium") +
  theme_minimal()
```



Al revisar la distribución de `potassium` por fabricante, notamos que el fabricante N tiene una media mucho más alta que los demás, sin embargo este nivel tiene solo 3 observaciones por lo que esta caja es poco representativa. El fabricante G presenta una varianza más chica que las demás, y es a su vez el que tiene más observaciones. El fabricante K tiene una media similar a la de G, pero con una varianza mucho más alta, lo que se debe a 4 valores muy altos (posibles valores atípicos) que podemos observar en el boxplot. La marcada asimetría derecha de las distribuciones por fabricante podría deberse a la existencia dentro de cada empresa de productos diferenciales, posiblemente suplementados o fortificados con minerales, con alto contenido de potasio como estrategia de venta. Los valores más altos de potasio se encuentran correlacionados

con los valores máximos de contenido de fibra, pudiendo ser esta suplementación con fibra la estrategia de marketing aplicada en los productos. En general, encontramos indicios de que puede haber diferencias significativas pero que se ven opacadas por el hecho de que para los niveles N, P, Q y R tenemos muy pocos datos y por lo tanto las estimaciones son mucho menos precisas. La distribución del contenido de potasio entre productos de los fabricantes R y G presenta mayor simetría. Generalmente el uso de gráficos de caja (box-plot) se desaconseja para casos en los que se cuenta con pocos datos, ya que las distintas medidas de la distribución resultan estimadas de manera muy precaria por los datos presentados.

Anova y supuestos

```
anova_model_potassium <- aov(potassium ~ mfr, data = data)
anova_result <- summary(anova_model_potassium)
print("ANOVA result for log(potassium) by manufacturer:")
```

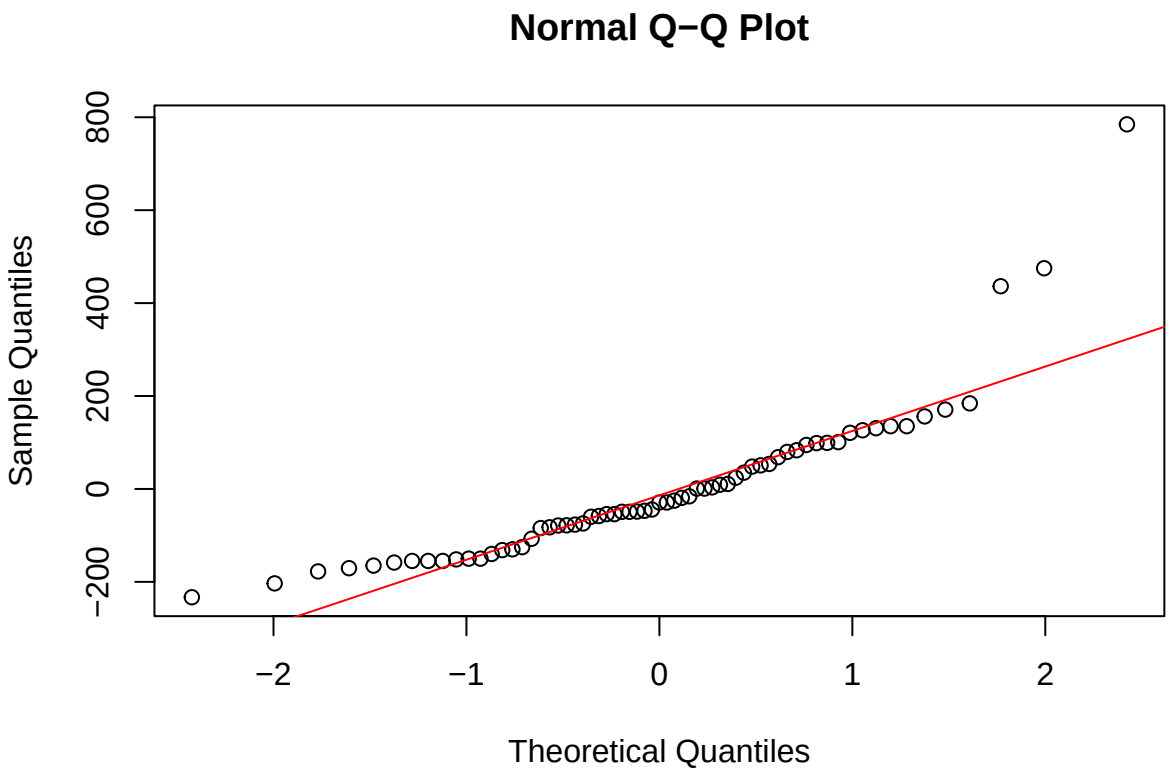
```
## [1] "ANOVA result for log(potassium) by manufacturer:"
```

```
print(anova_result)
```

```
##             Df  Sum Sq Mean Sq F value Pr(>F)
## mfr          5  315992   63198   2.113 0.0763 .
## Residuals   59 1764263   29903
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El valor del para el estadístico F es 2.113 y el valor p de la prueba F es 0.0763, entonces para un nivel de significancia del 0.05 no se rechaza la hipótesis nula, la evidencia no es lo suficientemente fuerte como para afirmar que hay diferencias significativas en las cantidades de potasio entre los diferentes fabricantes. Sin embargo, el valor p es menor a 0.10, lo que podría interpretarse como marginalmente significativo. Esto sugiere que existe una tendencia a que las diferencias entre grupos sean relevantes que posiblemente se ve afectada por el desbalance en los tamaño de muestra para cada nivel y en la presencia de valores atípicos. La falta de potencia dada por el bajo n para algunos de los niveles del factor fabricante pudo haber impedido que el modelo detecte diferencias entre los fabricantes, las cuales se pudieron apreciar en el análisis descriptivo.

```
# Gráfico QQ para verificar la normalidad de los residuos
qqnorm(residuals(anova_model_potassium))
qqline(residuals(anova_model_potassium), col = "red")
```




```
shapiro_test <- shapiro.test(residuals(anova_model_potassium))
print("Shapiro-Wilk normality test result:")
```

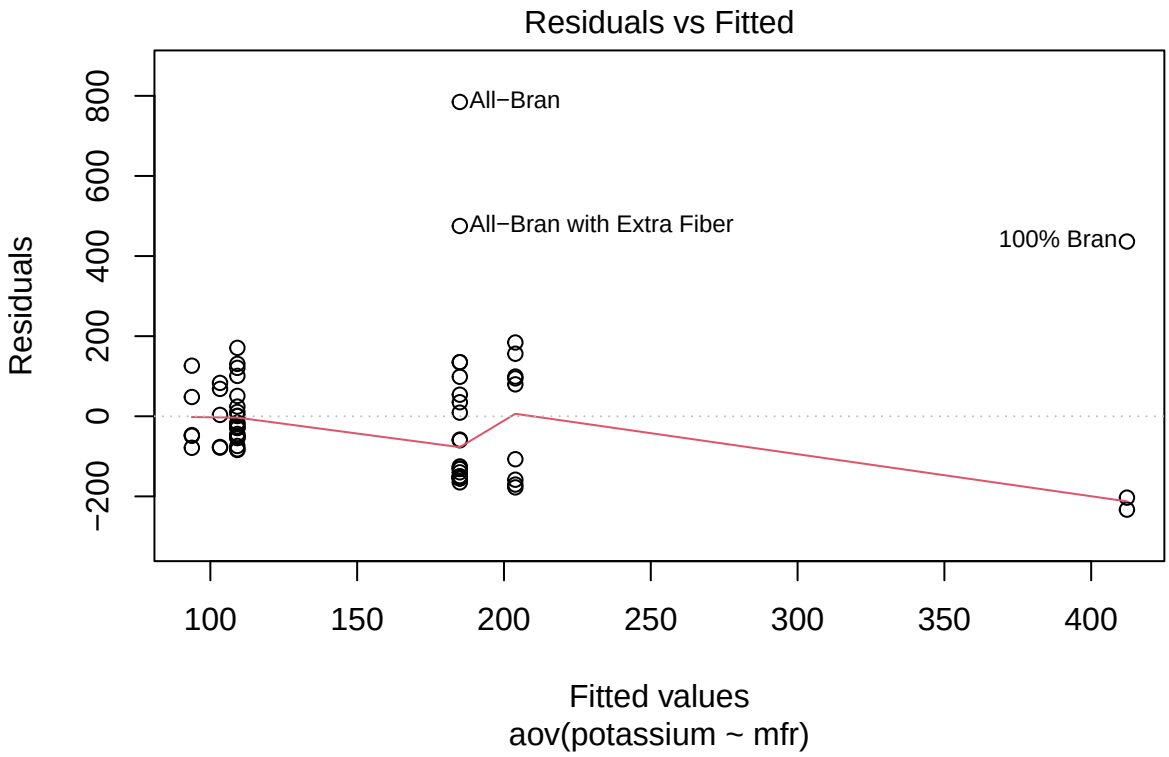
```
## [1] "Shapiro-Wilk normality test result:"
```

```
print(shapiro_test)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(anova_model_potassium)
## W = 0.82004, p-value = 1.911e-07
```

La prueba de normalidad de Shapiro-Wilk se utiliza para evaluar si una muestra proviene de una distribución normal. La hipótesis nula es que los datos se distribuyen normalmente. En este caso el valor de W es menor a 1 lo que sugiere que los residuos no son normales. El valor p de 1.911e-07 indica que la hipótesis nula es rechazada, bajo la suposición de que los datos son normales, la probabilidad de obtener un W tan bajo como 0.82004 es casi nula, por lo tanto no se cumple el supuesto de normalidad.

```
# Gráfico de caja para verificar la homogeneidad de varianzas
plot(anova_model_potassium, 1)
```




```
# Prueba de Levene para homogeneidad de varianzas
levene_test <- leveneTest(potassium ~ mfr, data = data)
print("Levene's test for homogeneity of variances result:")
```

```
## [1] "Levene's test for homogeneity of variances result:"
```

```
print(levene_test)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  5  1.3358 0.2618
##       59
```

El gráfico sugiere que la dispersión de los residuos no es homogénea entre niveles y que hay algunos valores atípicos. Por otro lado, el Test de Levene es una prueba estadística utilizada para evaluar la homogeneidad de varianzas entre dos o más grupos. La hipótesis nula es que las varianzas son iguales. En este caso, el valor p de 0.2618 indica que la hipótesis nula no se rechaza, lo que sugiere que no hay diferencias significativas en las varianzas entre los grupos. El hecho de que este es menos sensible a las desviaciones de la normalidad que otros tests (al permitir usar la mediana como centro) permite obtener un resultado más confiable sobre todo teniendo en cuenta que los residuos no son normales.

Estos resultados dan cuenta de que las conclusiones que extraigamos del modelo realizado no son válidas, puesto que no se cumple el supuesto de normalidad de los residuos. Para abordar estos problemas podemos aplicar una transformación a la variable dependiente. En este caso, probaremos con la transformación logarítmica, que es comúnmente utilizada para tratar datos que presentan una asimetría hacia la derecha,

como es el caso de nuestra variable `potassium`. Los datos asimétricos (sesgados a la derecha) tienen una cola larga en valores altos. Aplicar una transformación logarítmica suele "comprimir" esos valores altos, haciendo que la distribución se acerque más a la simetría.

Otra posible transformación, esta vez al factor, sería sumar los 4 niveles que tienen menor cantidad de datos (N,P,Q,R) y considerar un nuevo factor con 3 niveles. La nueva variable para fabricantes podría ser `G`, `K` y `Otros`. Esto podría ayudar a aumentar la potencia del modelo, al tener una estimación más acertada de la varianza; pero no podríamos sacar conclusiones individuales sobre los fabricantes agrupados en `Otros`.

En nuestro análisis también hemos identificado posibles outliers. Dado que estos datos forman parte de los valores reales de la variable `potassium` para los productos de cada fabricante y son datos valiosos para llegar a concluir si existen diferencia significativas entre ellos con respecto a dicha variable, consideramos que la forma de lidiar con ellos es a través de la transformación logarítmica, que tiende a reducir el impacto de los valores atípicos en el análisis, y que no deberían ser eliminados.

Por último dadas las características de la variable `potassium`, así como lo que hemos observado de las otras variables, podría considerarse utilizar modelos generalizados para realizar este análisis. Una aproximación básica de este último método se adjunta al final del ejercicio.

A continuación realizaremos la transformación logarítmica de la variable `potassium`, realizaremos los diferentes tests, también haremos el análisis agrupando la variable `mfr` en 3 niveles, y finalmente comentaremos los resultados.

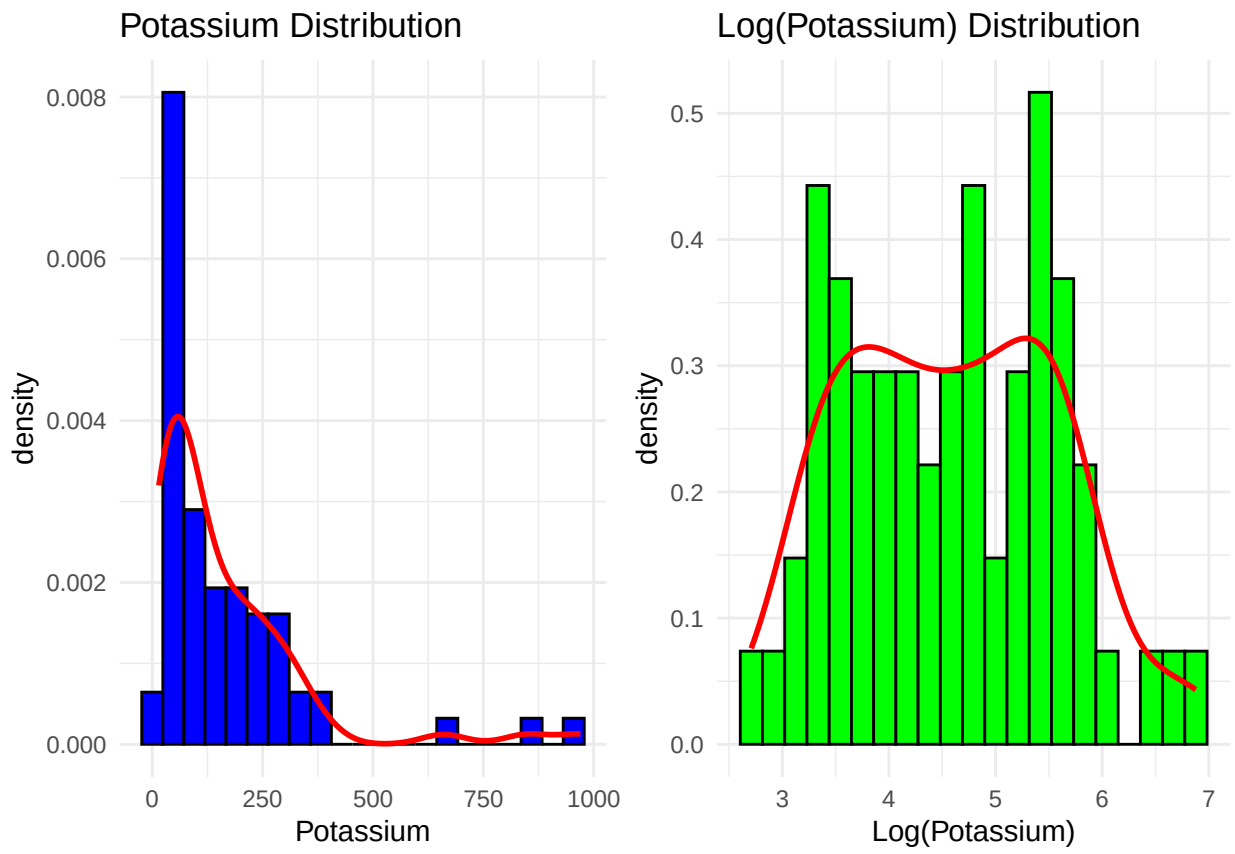
Transformación logarítmica

Al realizar la transformación de la variable respuesta, gráficamente se comprueba el cambio en la simetría de los datos. También es posible notar que estos aún no siguen una distribución normal, aunque ya es un avance respecto a la situación inicial.

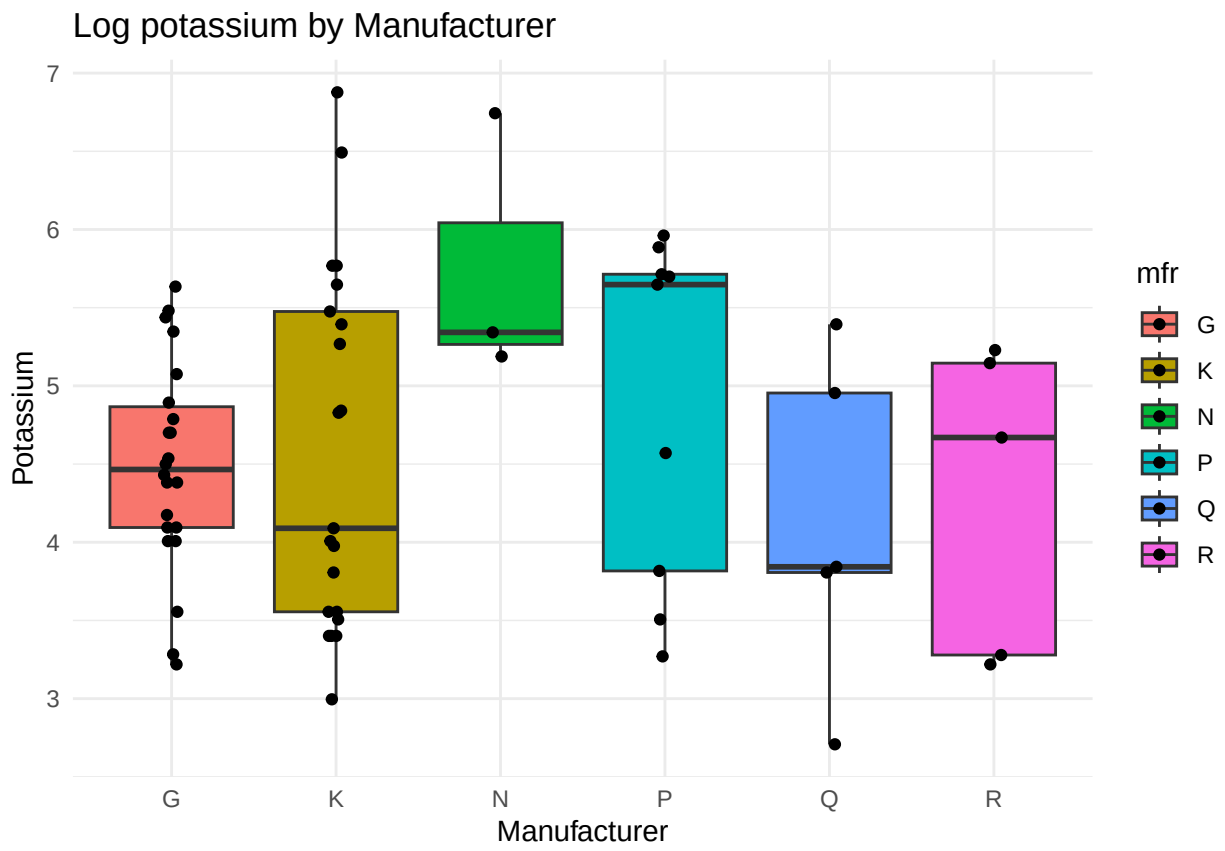
```
data$log_potassium <- log(data$potassium)
p1 <- ggplot(data, aes(x = potassium)) +
  geom_histogram(aes(y = after_stat(density)), fill="blue", binwidth = (max(data$potassium)-min(data$po
  geom_density(color="red", size=1) +
  ggtitle("Potassium Distribution") +
  xlab("Potassium") +
  theme_minimal()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
p2 <- ggplot(data, aes(x = log_potassium)) +
  geom_histogram(aes(y = after_stat(density)), fill="green", binwidth = (max(data$log_potassium)-min(da
  geom_density(color="red", size=1) +
  ggtitle("Log(Potassium) Distribution") +
  xlab("Log(Potassium)") +
  theme_minimal()

gridExtra::grid.arrange(p1, p2, ncol=2)
```




```r
# Distribución de la variable `log_potassium` por fabricante
ggplot(data, aes(x = mfr, y = log_potassium, fill = mfr)) +
  geom_boxplot() +
  geom_point(position = position_jitterdodge(jitter.width = 0.4)) +
  ggtitle("Log potassium by Manufacturer") +
  xlab("Manufacturer") +
  ylab("Potassium") +
  theme_minimal()
```



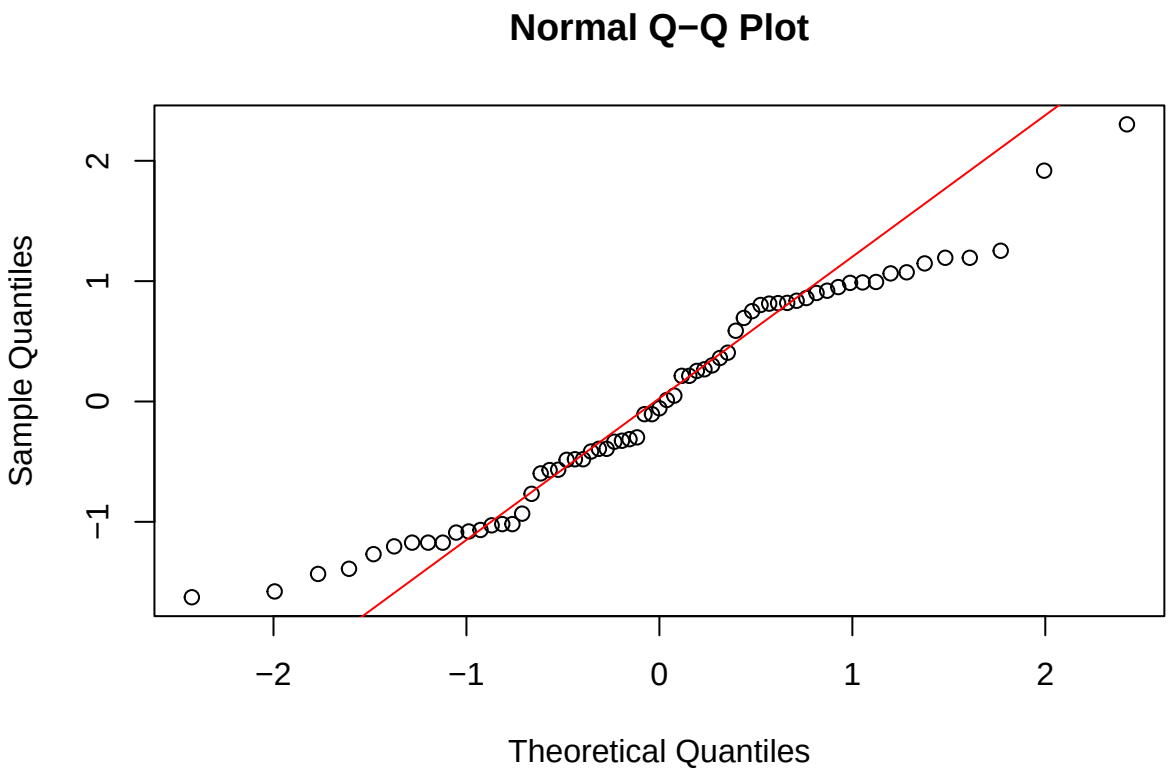

```
# ANOVA para log(potassium) por fabricante
anova_model_log_potassium <- aov(log_potassium ~ mfr, data = data)
anova_result <- summary(anova_model_log_potassium)
print("ANOVA result for log(potassium) by manufacturer:")
```

```
## [1] "ANOVA result for log(potassium) by manufacturer:"
```

```
print(anova_result)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## mfr          5   6.58  1.3159   1.396  0.239
## Residuals   59  55.60  0.9424
```

```
# Gráfico QQ para verificar la normalidad de los residuos
qqnorm(residuals(anova_model_log_potassium))
qqline(residuals(anova_model_log_potassium), col = "red")
```



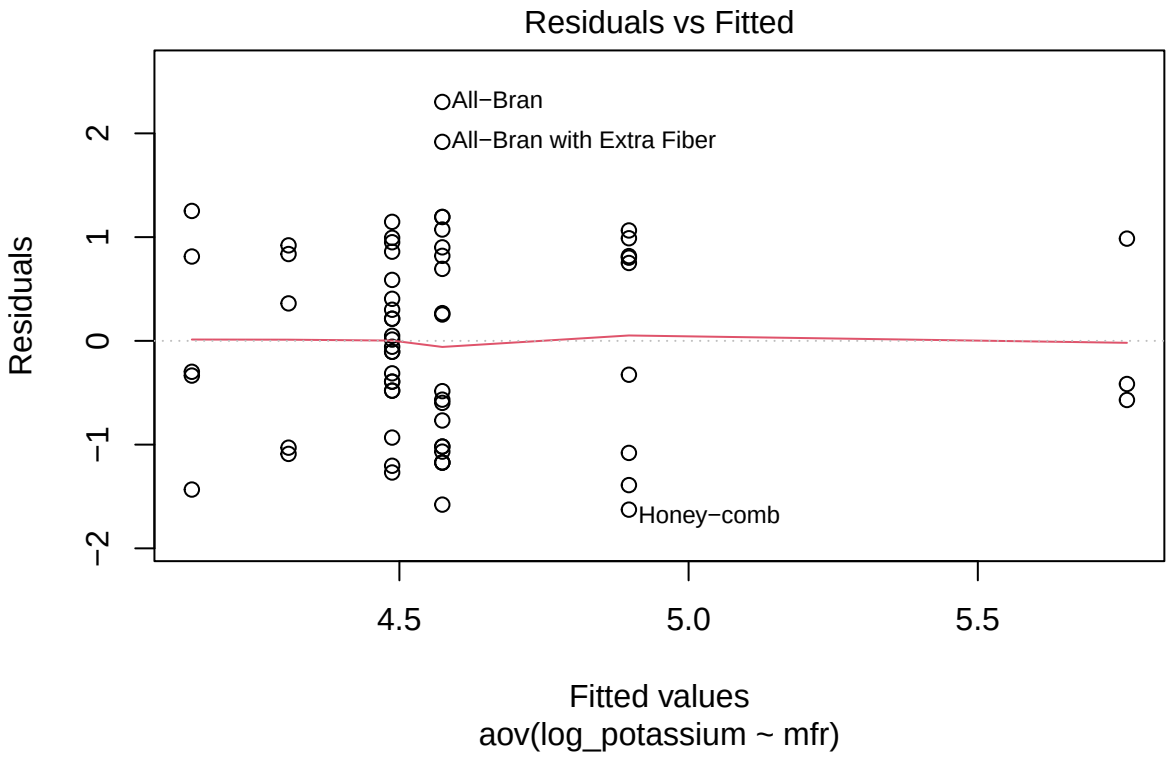

```
shapiro_test <- shapiro.test(residuals(anova_model_log_potassium))
print("Shapiro-Wilk normality test result:")
```

```
## [1] "Shapiro-Wilk normality test result:"
```

```
print(shapiro_test)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(anova_model_log_potassium)
## W = 0.95983, p-value = 0.03361
```

```
# Gráfico de caja para verificar la homogeneidad de varianzas
plot(anova_model_log_potassium, 1)
```




```
# Prueba de Levene para homogeneidad de varianzas
levene_test <- leveneTest(potassium ~ mfr, data = data)
print("Levene's test for homogeneity of variances result:")
```

```
## [1] "Levene's test for homogeneity of variances result:"
```

```
print(levene_test)
```

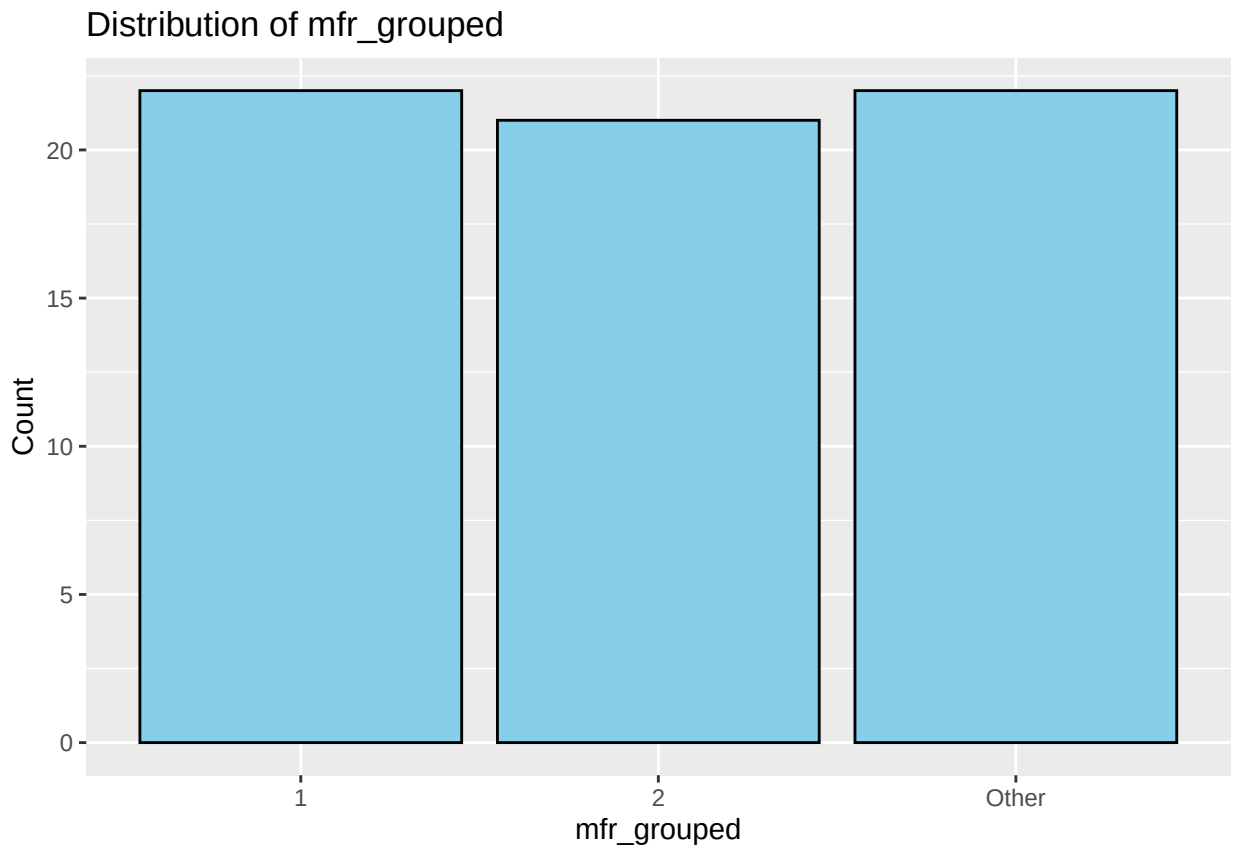
```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  5  1.3358 0.2618
##       59
```

Es posible notar que al aplicar la transformación de la variable respuesta no se ha solucionado plenamente el incumplimiento del supuesto de normalidad. La variable presenta una distribución notoriamente alejada de la normal en las colas y muy irregular en el centro, haciendo que el test de normalidad de Shapiro-Wilks mantenga su status anterior de significativo. Por otro lado, la transformación tampoco modificó la conclusión a tomar respecto al impacto del fabricante sobre el contenido de potasio. A esto se debe considerar que cualquier conclusión sobre este modelo hubiera tenido que considerar que la variable respuesta del modelo ahora se trataba de una transformación de la variable original, por lo que las conclusiones a tomar no se podrían haber extrapolado de manera directa a ella. En base a lo explicado y a los resultados obtenidos, una transformación no constituye una alternativa adecuada para el análisis de estos datos.

Factor agrupado

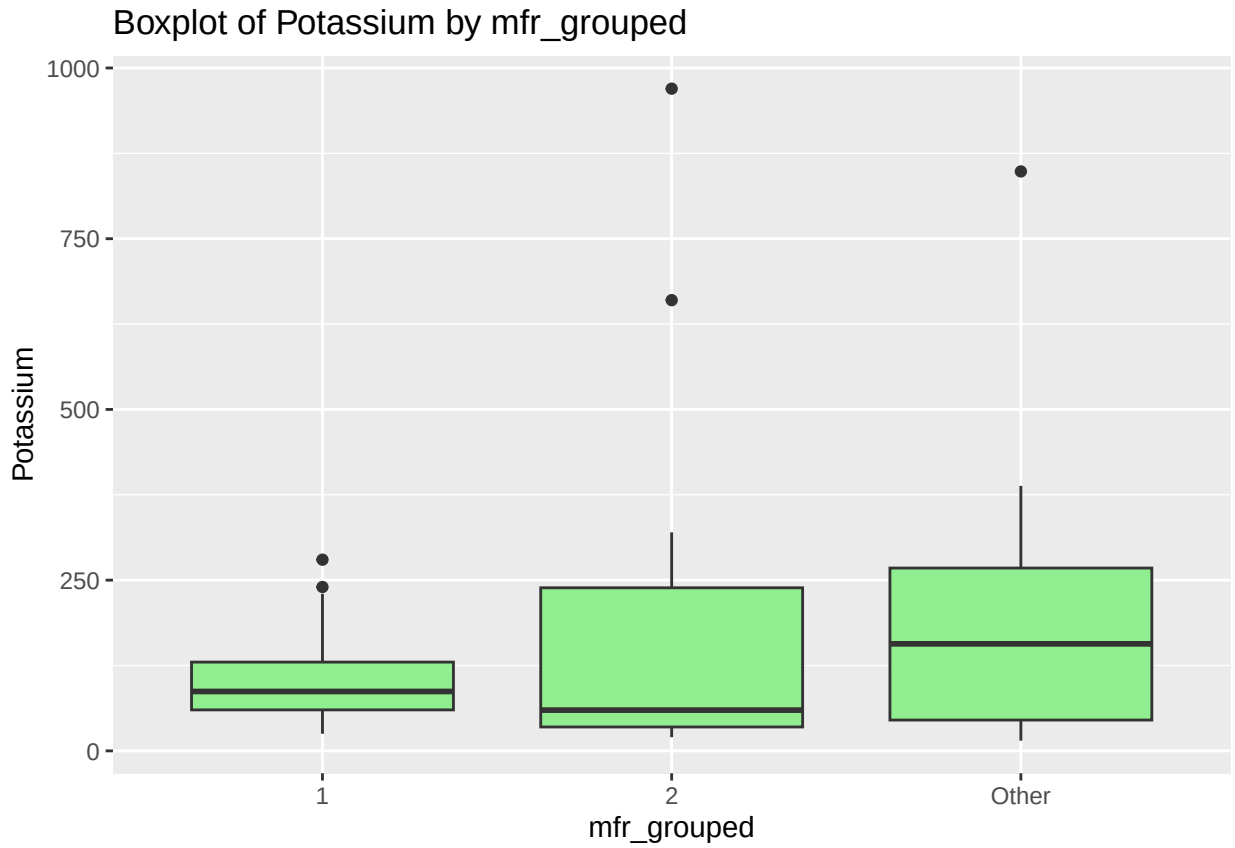
Con el objeto de aumentar la potencia del modelo, se decidió agrupar las empresas con pocos productos en una única categoría "Other". Gráficamente se observa que ahora el factor fabricante presenta tres niveles con n más parejos, mostrándose además más armonioso el gráfico de cajas.

```
# Agrupando los niveles de mfr
data <- data %>%
  mutate(mfr_grouped = ifelse(mfr %in% c('G', 'K'), mfr, 'Other'))

data$mfr_grouped <- as.factor(data$mfr_grouped)

# Graficamos la distribución de la nueva variable categórica
p_dist <- ggplot(data, aes(x = mfr_grouped)) +
  geom_bar(fill = 'skyblue', color = 'black') +
  ggtitle('Distribution of mfr_grouped') +
  xlab('mfr_grouped') +
  ylab('Count')
p_dist
```




```
# Boxplot de potassium por mfr_grouped
p_box <- ggplot(data, aes(x = mfr_grouped, y = potassium)) +
  geom_boxplot(fill = 'lightgreen') +
  ggtitle('Boxplot of Potassium by mfr_grouped') +
  xlab('mfr_grouped') +
```

```
  ylab('Potassium')
p_box
```




```
# Modelo
anova_model <- aov(potassium ~ mfr_grouped, data = data)
summary(anova_model)
```

```
##             Df  Sum Sq Mean Sq F value Pr(>F)
## mfr_grouped  2   82887   41443   1.286  0.284
## Residuals   62 1997368   32216
```

```
# Test de normalidad de Shapiro-Wilk
shapiro_test <- shapiro.test(residuals(anova_model))
print(shapiro_test)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(anova_model)
## W = 0.75266, p-value = 4.049e-09
```

```
# Test de homogeneidad de varianzas de Levene
levene_test <- leveneTest(potassium ~ mfr_grouped, data = data)
print(levene_test)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  2.3192 0.1068
##       62
```

Esta nueva estrategia no logró mejorar el ajuste del modelo, ni favoreció que el mismo cumpla el supuesto de normalidad de los errores. A partir de esto, creemos necesario probar de ingresar otros factores al modelo para intentar que el mismo sea interpretable.

d)

Otra variable que podría considerarse como factor es `shelf`, debido entre otros factores a que presenta una distribución más homogénea que las otras variables categóricas. Además, de acuerdo con la suposición de que es probable de que haya sido una variable controlada en la realización del estudio, tendría más sentido considerarla como un factor fijo. Por último los estudios con esta variable podrían darnos información valiosa acerca de si hay algún tipo de relación entre la posición del producto en la góndola y su contenido nutricional. Por ejemplo, si consideramos la variable `sugars` como variable respuesta, podríamos analizar si los cereales con más azúcares tienden a estar en estantes más altos (más visibles) que los cereales con menos azúcares. Esta relación podría ser de interés para identificar cómo se posicionan estratégicamente productos de diferentes características nutricionales ante los ojos de los clientes, estando los cereales de estantes más altos a la altura de la visión del consumidor, mientras los cereales del estante de más abajo pueden ser fácilmente ignorados.

ANEXO 1 Cálculos para completar tabla ANOVA

```r
# Calcular la media global
media_global_potassium <- mean(data$potassium)


# Suma de cuadrados total (SST)
SST <- sum((data$potassium - media_global_potassium)^2)

# Calcular las medias por tratamiento y el tamaño de cada grupo
medias_trat <- tapply(data$potassium, data$mfr, mean)
n_j <- tapply(data$potassium, data$mfr, length)
# Suma de cuadrados del tratamiento (SSTR)
# CADA TRATAMIENTO SE ESCALA POR EL PESO CORRESPONDIENTE!
SSTr <- sum(n_j * (medias_trat - media_global_potassium)^2)

# Suma de cuadrados del error (SSE)
SSE <- SST - SSTr

# Cuadrado medio de los tratamientos
MSTr = SSTr / (length(unique(data$mfr)) - 1)

# Cuadrado medio del error
# I(J-1) = total de observaciones menos la cantidad de niveles de la variable
MSE = SSE / (length(data$potassium) - length(unique(data$mfr)))
```

```r
# Imprimir los resultados
cat("SST (Suma de Cuadrados Total):", SST, "\n")
```

```
## SST (Suma de Cuadrados Total): 2080254
```

```r
cat("SSTR (Suma de Cuadrados del Tratamiento):", SSTr, "\n")
```

```
## SSTR (Suma de Cuadrados del Tratamiento): 315991.7
```

```r
cat("SSE (Suma de Cuadrados del Error):", SSE, "\n")
```

```
## SSE (Suma de Cuadrados del Error): 1764263
```

```r
cat("MSTr (Cuadrado Medio de los Tratamientos):", MSTr, "\n")
```

```
## MSTr (Cuadrado Medio de los Tratamientos): 63198.34
```

```r
cat("MSE (Cuadrado Medio del Error):", MSE, "\n")
```

```
## MSE (Cuadrado Medio del Error): 29902.76
```

ANEXO 2 Análisis adicionales: modelo generalizado

En el marco de la Maestría en Estadística Aplicada (MAEA, FCE-UNC) hemos empezado a aplicar el ajuste de modelos lineales generalizados. Dado que en este caso no hemos logrado ajustar un modelo para las variables de interés, en este módulo se intentará una aproximación de ajuste mediante la vía de los GLM. Dadas las características de la variable respuesta (continua, asimétrica), se optó por emplear la familia Gamma, utilizando la función de enlace canónica para dicha distribución (inversa).

```r
generalized <- glm(potassium ~ mfr, data, family = "Gamma")
summary(generalized)
```

```
##
## Call:
## glm(formula = potassium ~ mfr, family = "Gamma", data = data)
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0091578  0.0018665   4.906 7.65e-06 ***
## mfrK        -0.0037512  0.0021808  -1.720  0.09066 .
## mfrN        -0.0067316  0.0022971  -2.930  0.00481 **
## mfrP        -0.0042528  0.0024345  -1.747  0.08586 .
## mfrQ         0.0015156  0.0049301   0.307  0.75961
## mfrR         0.0005242  0.0045406   0.115  0.90848
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.9138874)
```

```
##
##     Null deviance: 62.310  on 64  degrees of freedom
## Residual deviance: 52.618  on 59  degrees of freedom
## AIC: 790.07
##
## Number of Fisher Scoring iterations: 7
```

Por defecto la función glm() emplea la función de enlace canónica cuando no se realiza ninguna aclaración.

El modelo obtenido se puede comparar mediante el criterio de Akaike (AIC) con el modelo lineal general obtenido en pasos anteriores, dado que ambos modelos contienen las mismas variables respuesta y factor.

```
linear <- lm(potassium ~ mfr, data)

AIC(linear)
```

```
## [1] 862.0377
```

```
AIC(generalized)
```

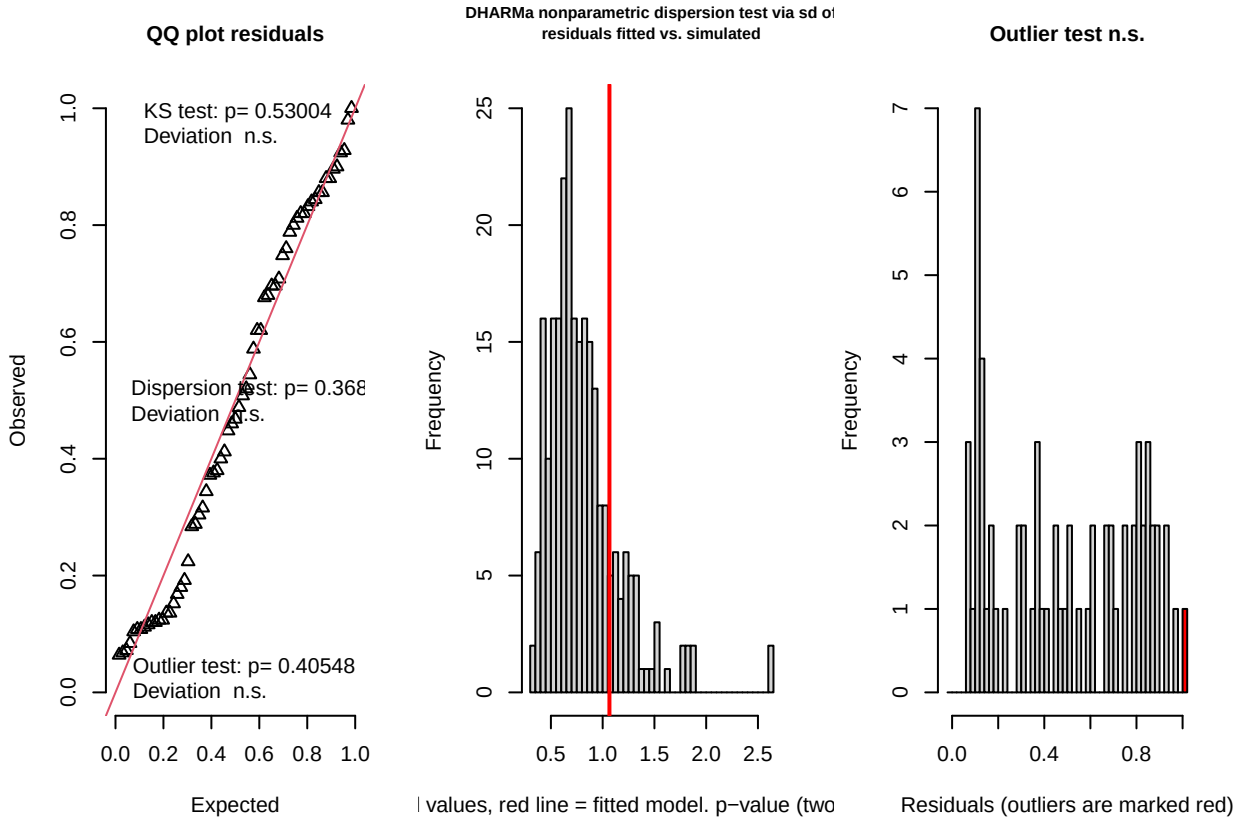
```
## [1] 790.0748
```

El AIC del modelo generalizado es menor (790.08 vs. 862.04), indicando mejor ajuste. De arranque, es preferible.

```
library(DHARMa)
```

```
## This is DHARMa 0.4.7. For overview type '?DHARMa'. For recent changes, type news(package = 'DHARMa')
```

```
DHARMa::testResiduals(generalized)
```




```
## $uniformity
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  simulationOutput$scaledResiduals
## D = 0.10031, p-value = 0.53
## alternative hypothesis: two-sided
##
##
## $dispersion
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 1.3031, p-value = 0.368
## alternative hypothesis: two.sided
##
##
## $outliers
##
##  DHARMa outlier test based on exact binomial test with approximate
##  expectations
##
## data:  simulationOutput
## outliers at both margin(s) = 1, observations = 65, p-value = 0.4055
```

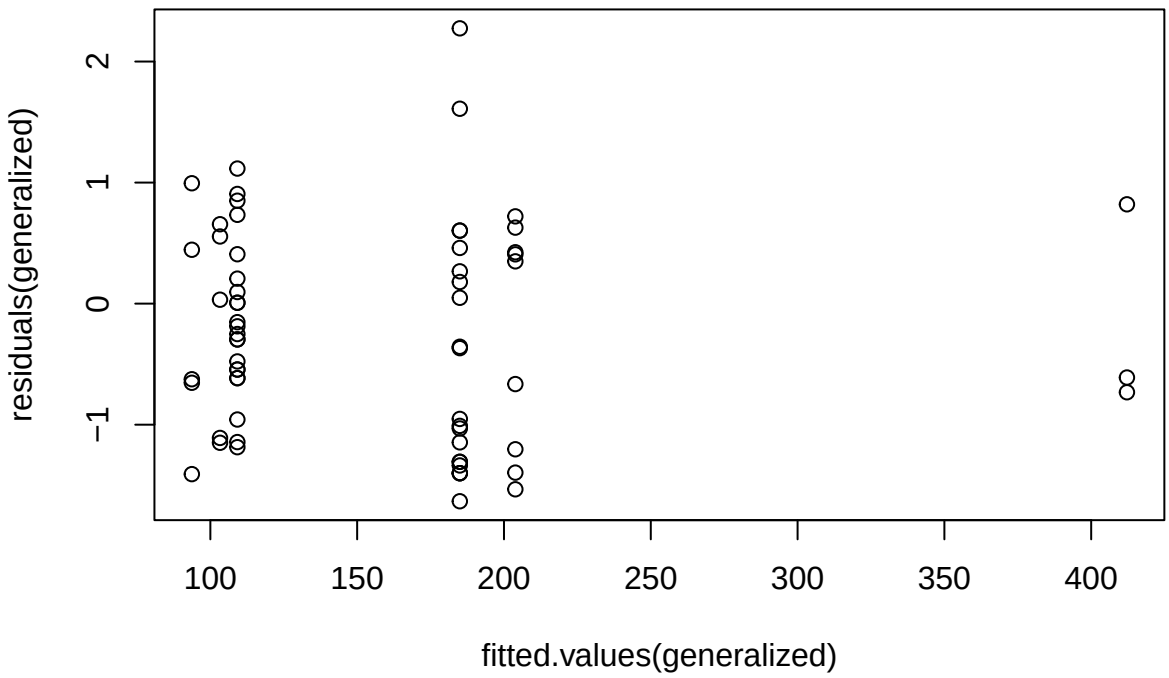
```
## alternative hypothesis: true probability of success is not equal to 0.007968127
## 95 percent confidence interval:
##  0.0003894289 0.0827630877
## sample estimates:
## frequency of outliers (expected: 0.00796812749003984 )
##                                             0.01538462
```

```
res <- DHARMa::simulateResiduals(generalized)
testResiduals(res)
```

```
## $uniformity
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  simulationOutput$scaledResiduals
## D = 0.10031, p-value = 0.53
## alternative hypothesis: two-sided
##
##
## $dispersion
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 1.3031, p-value = 0.368
## alternative hypothesis: two.sided
##
##
## $outliers
##
##  DHARMa outlier test based on exact binomial test with approximate
##  expectations
##
## data:  simulationOutput
## outliers at both margin(s) = 1, observations = 65, p-value = 0.4055
## alternative hypothesis: true probability of success is not equal to 0.007968127
## 95 percent confidence interval:
##  0.0003894289 0.0827630877
## sample estimates:
## frequency of outliers (expected: 0.00796812749003984 )
##                                             0.01538462
```

Para este modelo generalizado, los test de la librería especializada DHARMa no detectan exceso de valores extremos o desviaciones significativas de la normal. Adicionalmente, no se observa heterocedasticidad ni presencia de subdispersión o sobredispersión.

```
plot(residuals(generalized) ~ fitted.values(generalized))
```



Retomando los valores obtenidos en el resumen del modelo generalizado, recordando que el Intercept en estos modelos representa la media esperada para el primer nivel del factor (en este caso el fabricante G), se concluye que se presentaron diferencias significativas en el contenido de potasio de los cereales de General Mill's (G) y Nabisco (N) ($p < 0{,}05$).

```
library(multcomp)
```

```
## Loading required package: mvtnorm

## Loading required package: survival

## Loading required package: TH.data

##
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':
##
##     geyser
```

```
test.tukey <- glht(generalized, linfct = mcp(mfr="Tukey"))
cld(test.tukey)
```

```
##    G    K    N    P    Q    R
##  "a" "ab"  "b" "ab" "ab" "ab"
```