Actividad 1: Análisis Previo de Datos

Ejercicio 1: Gráficos

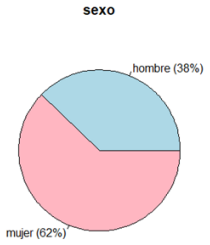


Figura 1. Gráfico de tarta por sexo

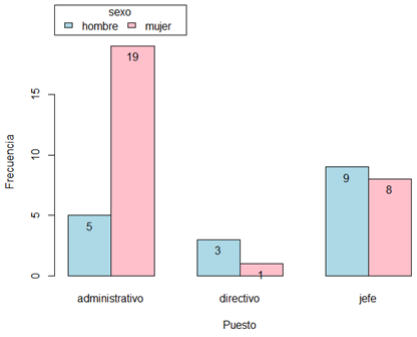


Figura 2. Gráfico de barras agrupadas Puesto-Sexo

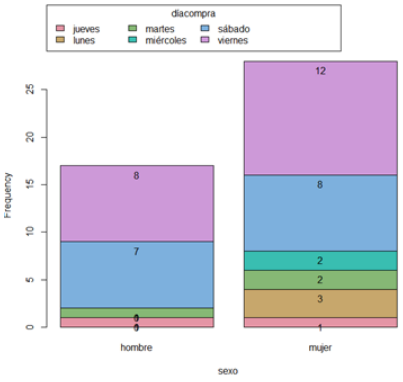


Figura 3. Gráfico de barras agrupadas Sexo-Día compra

Podemos observar que en este estudio hay un **porcentaje mayor de mujeres**. En cuento a puestos de trabajo, la posición más baja cuenta con un alto número de mujeres en comparación de hombres, **ocupando los hombres más puestos en altos cargos**. La mayoría de personas en este estudio optan por realizar la **compra viernes y sábado**, aunque las mujeres, al contrario que los hombres, tienden a repartírsela algo más durante la semana.

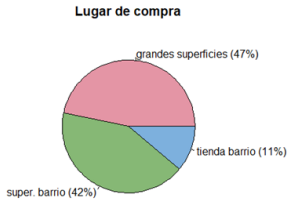


Figura 4. Gráfico de sectores por lugar de compra

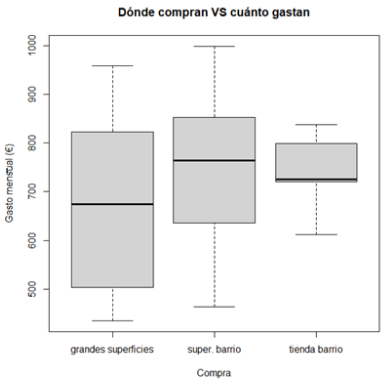


Figura 5. Boxplot Lugar de compra - Gatos mensual

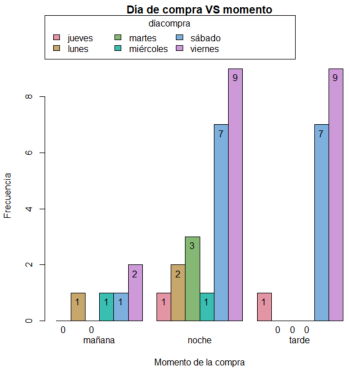


Figura 6. Gráfico de barras Compras-Momento

Luego, en cuanto a las compras, hay una **tendencia a realizarlas los días laborables por la noche**, aunque el día preferido sea el sábado por la tarde-noche. La gente en este estudio prefiere ligeramente **comprar en grandes superficies**, casi empatando con las personas que eligen comprar en el supermercado de barrio. Los que eligen comprar en la tienda de barrio son una alta minoría (11%). Los rangos de gastos mensuales por persona son más altos para los que compran en supermercados de barrio. En cambio, **en las grandes superficies el rango de precios es más equilibrado**, pues encontramos a los que menos gastan mensualmente. Los que optan por comprar en tiendas de barrio son aquellos que, en general, gastan significativamente más de manera mensual.

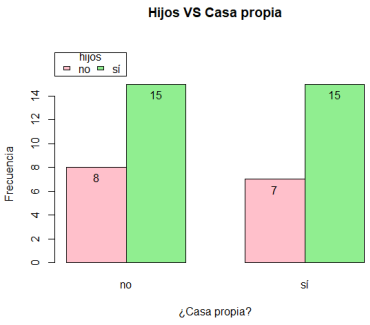


Figura 7. Diagrama de barras Hijos-Casa propia

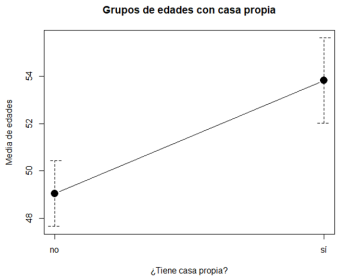


Figura 8. Gráfico de medias Casa propia-Media de edades

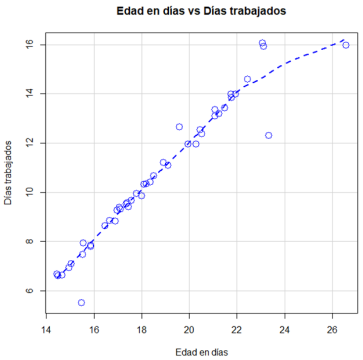


Figura 9. Diagrama de dispersión Edad-Días trabajados

Vemos en Figura 7 que **tener casa propia y tener hijos no está correlacionado**, pues solo la mitad de la muestra cumple ambas condiciones. Además, **los que tienen casa propia suelen ser los que superan los 52 años**. Lo que sí está correlacionado es la edad con los días trabajados, lo que demuestra que **cuanto más mayor seas, más habrás trabajado**.

Ejercicio 2: Datos ausentes

1) ¿Hay datos ausentes? No.

```
> sapply(datos0, function(x)(sum(is.na(x))))
       DNI       tfno     ciudad       sexo      hijos  casapropia     compra 
         0          0          0          0          0          0          0 
horacompra gastomensual  díacompra fecha_nac      edad  compra_A edadendías 
         0          0          0          0          0          0          0 
    puesto     sueldo fecha_trab años_trab      días hora_cod 
         0          0          0          0          0          0 
```

Figura 10. Código en R de búsqueda de datos ausentes (NAs)

> Por lo tanto, no podemos continuar con el ejercicio.

En el caso de que pudiéramos:

1. Localizamos los registros que tienen datos faltantes → which(is.na(datos[,1])) o vis_miss (datos) del paquete vis_dat
2. Para **saber si la pérdida de datos es aleatoria o sigue un patrón**, tomamos 2 variables con valores nulos (se asume que 1 dependiente, 1 independiente). Hacemos **test de varianzas** para saber si son iguales
3. Si lo son, hacemos **t test** para ver si las **medias** son iguales. (También podemos optar por ver en una matriz de correlaciones si están correlacionadas) → rstatix cor_matr

Si se cumplieran igualdad de varianzas y medias, quiere decir que las variables son dependientes y, por lo tanto, los valores ausentes siguen un patrón.

En cuanto al **tratamiento de estos valores ausentes**, si no hay patrones, podemos optar por eliminar los registros con datos ausentes (na.omit). En cambio, si hay dependencia entre variables con datos ausentes, podemos optar por rellenar los datos por la media, por regresión lineal o por constantes.

Ejercicio 3: Datos atípicos

Realizamos **diagramas de caja** para localizar valores atípicos de variables numéricas:

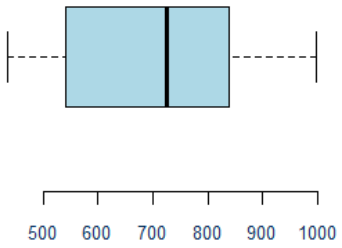


Figura 11. Boxplot gastomensual

No hay valores atípicos

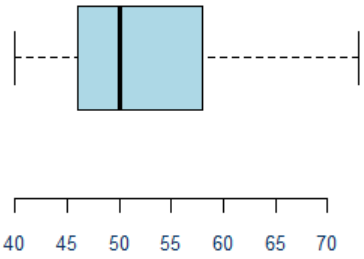


Figura 12. Boxplot edad

No hay valores atípicos

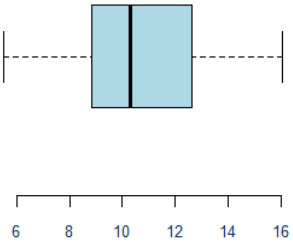


Figura 13. Boxplot días (trabajados)

No hay valores atípicos

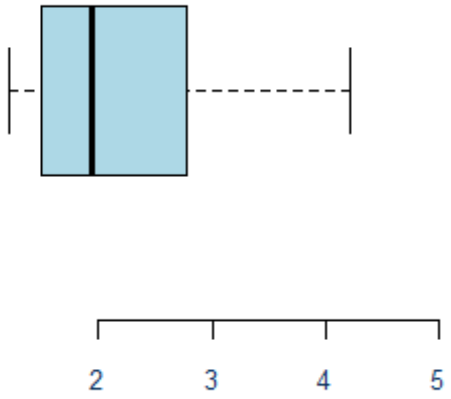


Figura 14. Boxplot sueldo

Por encima de 4.250€: 1 dato

5.690 €

Profesión: Directivo

Sexo: Hombre

Experiencia: 15 años

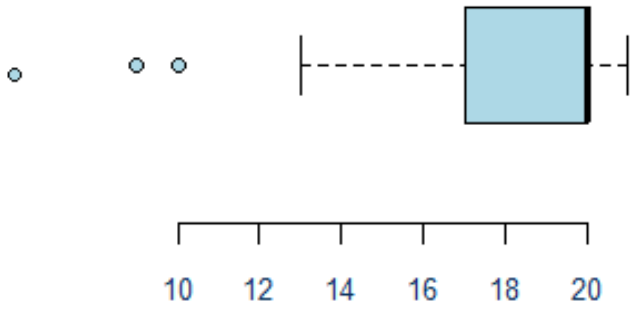


Figura 15. Boxplot horacompra

Por debajo de las 13h: 4 datos

9h (3 datos) 10h (1 dato)

75% compra en día laborable

75% compra en grandes superficies

75% hombres

Para el **tratamiento** de estos datos atípicos, podemos:

- Eliminarlos
- Reemplazarlos

- Transformarlos

Para variable horacompra: En este caso, para no perder el resto de datos de estas personas, reemplazaremos los datos de la hora de compras por las 13h para tampoco perder la información de que realizan la compra por la mañana, y así nos deshacemos de valores atípicos:

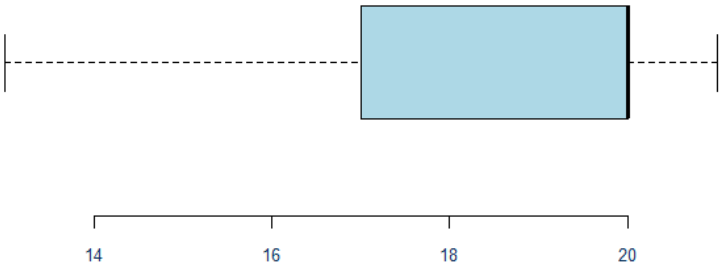


Figura 16. Boxplot de horacompra tras eliminar valores atípicos

Para variable sueldo: Haremos lo mismo, sustituiremos el sueldo por el valor más alto dentro del rango intercuartílico (4250€). Así, del mismo modo, conservaremos la información de que la persona tiene el mayor sueldo y a la vez nos deshacemos de valores atípicos:

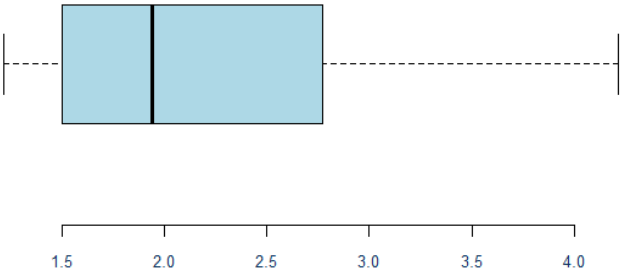


Figura 17. Boxplot de sueldo tras eliminar valores atípicos

Ejercicio 4: Verificación de supuestos básicos

1. **Normalidad**

   Hacemos uso del **test Shapiro-Wilk** para comprobar si la variable sigue una distribución normal y obtenemos el histograma de cada variable numérica y su Q-Q Plot

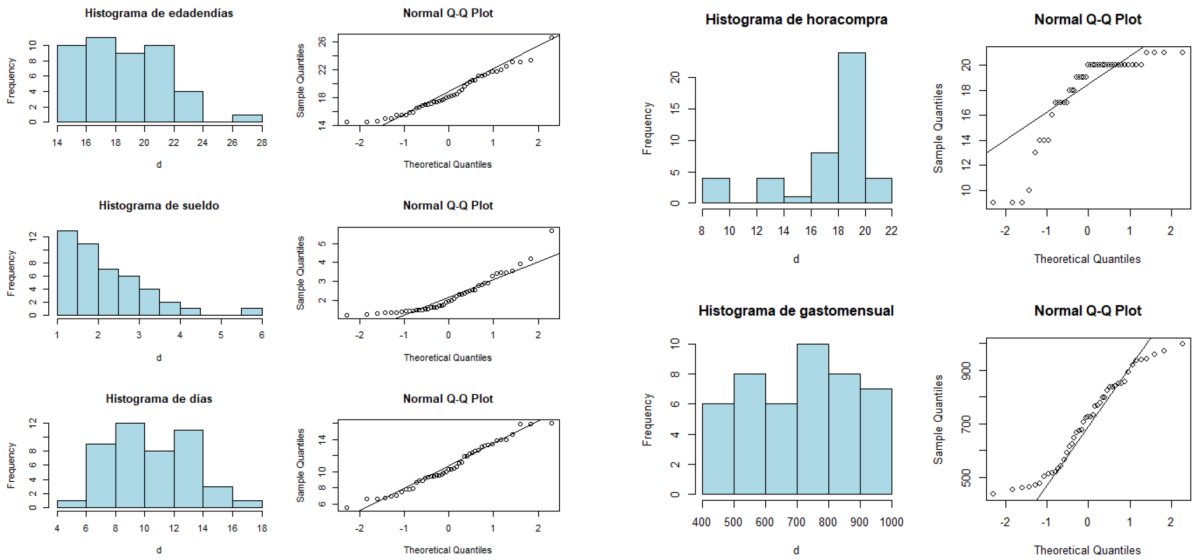


Figura 18. Gráficos de variables numéricas

Según la variable, ¿tenemos suficiente evidencia para afirmar que sigue una distribución normal?:

- ciudad: p-valor < 0.05 → NO
- edadendías: p-valor > 0.05 → SÍ
- sueldo: p-valor < 0.05 → NO
- días: p-valor > 0.05 → SÍ
- horacompra: p-valor < 0.05 → NO
- gastomensual: p-valor <0.05 → NO

2. **Linealidad y dependencia**

   Observamos la matriz de correlaciones para las variables numéricas:

	ciudad	horacompra	gastomensual	edadendías	sueldo	días
ciudad	1.00000000	0.168750808	0.276018961	0.14790710	-0.085857210	0.2238177
horacompra	0.16875081	1.000000000	0.191568514	-0.15511083	0.003003559	-0.1224293
gastomensual	0.27601896	0.191568514	1.000000000	0.28471920	-0.001461943	0.2792513
edadendías	0.14790710	-0.155110833	0.284719200	1.00000000	-0.099590782	0.9663766
sueldo	-0.08585721	0.003003559	-0.001461943	-0.09959078	1.000000000	-0.1496392
días	0.22381765	-0.122429262	0.279251323	0.96637662	-0.149639161	1.0000000

Figura 19. Matriz de correlaciones de las variables numéricas

De esta, podemos concluir que no existe relación lineal entre las variables, pues la correlación es muy baja, excepto:

- **Días (trabajados) y la edad en días**. Como la correlación es cercana a 1, podemos decir que existe una relación lineal fuerte entre las dos variables. Cuando una aumenta, la otra también tiende a aumentar. Esto también lo comprobamos anteriormente con un diagrama de dispersión (Figura 9), de lo que concluimos que es lógico ya que **cuanto más hayamos trabajado, mayores seremos**.

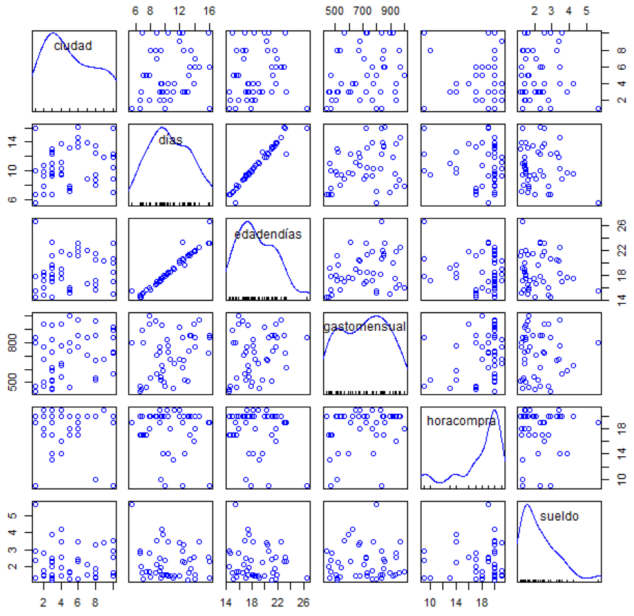


Figura 20. Matriz de diagramas de dispersión de las variables numéricas

Atendiendo a la Figura 20, podemos ver la clara relación lineal entre las variables días y edadendías. No es así para las demás variables, que se encuentran dispersas sin un patrón lineal.

Para el caso de variables categóricas, realizamos la prueba de **Chi-cuadrado**. Al revisar los p-valores de las pruebas de variables 2 a 2, observamos que **puede existir dependencia entre las variables**:

- `díacompra` y `compra_A` (p-valor < 0.05)

Esto tiene sentido ya que los días de compra se organizan por L,M,X,J,V,S,D y la variable compra_A también indica el día en el que se compra, solo que lo organiza en días laborables y fines de semana.

Las demás variables categóticas se asumen como independientes.

3. **Homocedastidad**

   Ajustamos modelos de regresión lineal:

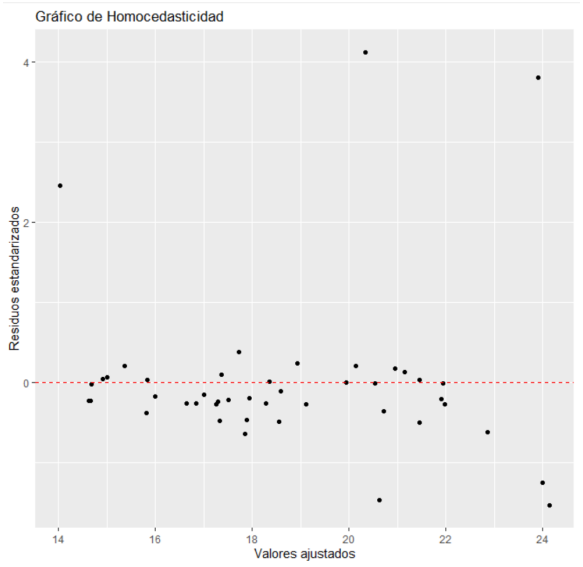
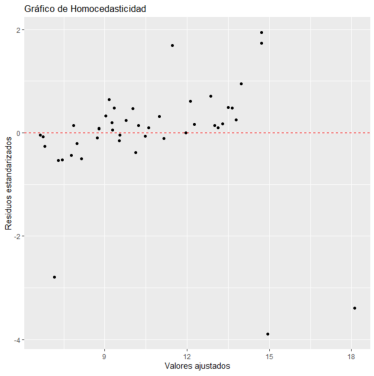


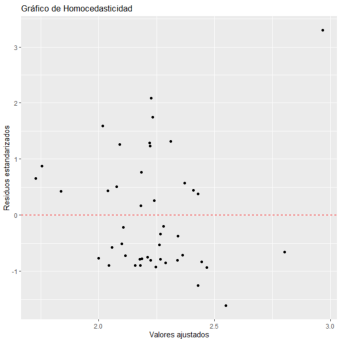
Figura 21. Gráfico de Homocedastidad de variables numéricas

`(edadendías ~ gastomensual + días + sueldo)`

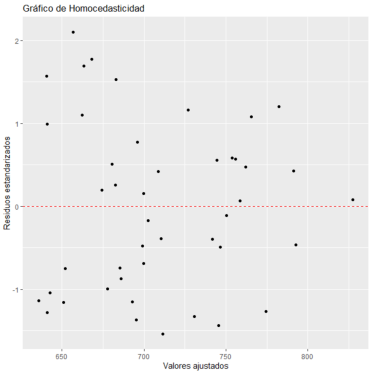
Haciendo la prueba de **Breusch-Pagan** (bptest de lmtest) , obtenemos un p-valor > 0.05, lo que quiere decir que hay suficiente evidencia para afirmar que **la homocedasticidad está presente en el modelo**. Esto significa que los errores (o **desviaciones**) son **iguales** para todas las observaciones



(días~ edadendías + gastomensual + sueldo)



(sueldo ~ edadendías + días + gastomensual)



(gastomensual ~ edadendías + días + sueldo)

En cambio, tenemos suficiente evidencia para afirmar que en estos modelos **NO se cumple la homocedasticidad**, pues p-valor < 0.05. Esto quiere decir que la **varianza NO es constante**, por lo que nuestro modelo pierde potencia estadística y los graficos no siguen ese patrón horizontal. La causa puede ser la mala elección del modelo.