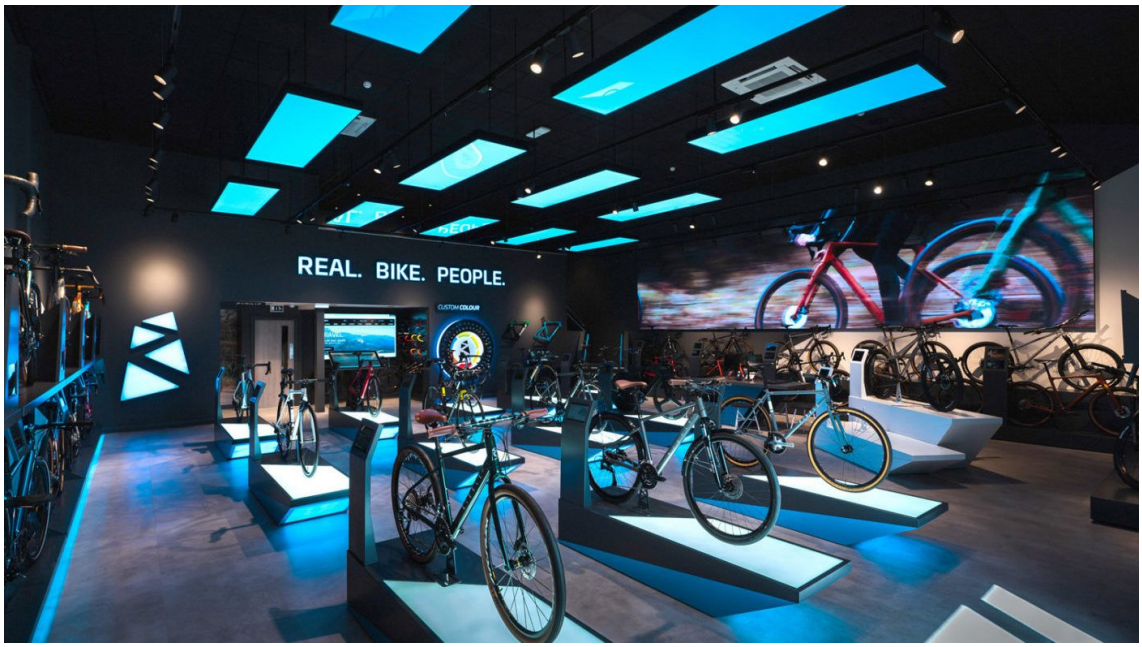
Análisis exploratorio y modelos de aprendizaje:

Caso de estudio Adventure Works



Adrià Sagrera

ÍNDICE

1. ANÁLISIS EXPLORATORIO DE LOS DATOS (EDA)

Figura 1 Comandos básicos R

```
> str(Clientes)
tibble [18,484 × 19] (S3: tbl_df/tbl/data.frame)
 $ TotalAmount      : num [1:18484] 8139 2994 4118 4631 3400 ...
 $ BikePurchase     : num [1:18484] 1 1 1 1 1 1 1 1 1 1 ...
 $ CustomerID       : num [1:18484] 11003 14501 21768 25863 28389 ...
 $ Country          : chr [1:18484] "Australia" "Southwest" "Canada" "Northwest" ...
 $ CountryRegionCode: chr [1:18484] "AU" "US" "CA" "US" ...
 $ Group            : chr [1:18484] "Pacific" "North America" "North America" "North America"
...
 $ PersonID         : num [1:18484] 11358 11211 14078 10553 15519 ...
 $ PersonType       : chr [1:18484] "IN" "IN" "IN" "IN" ...
 $ DateFirstPurchase: POSIXct[1:18484], format: "2001-07-01" "2001-07-01" ...
 $ BirthDate        : POSIXct[1:18484], format: "1968-02-15" "1938-05-13" ...
 $ Age              : num [1:18484] 52 82 74 74 56 55 57 48 59 66 ...
 $ MaritalStatus    : chr [1:18484] "S" "M" "S" "S" ...
 $ YearlyIncome     : chr [1:18484] "50001-75000" "75001-100000" "50001-75000" "25001-50000"
...
 $ Gender           : chr [1:18484] "F" "M" "M" "F" ...
 $ TotalChildren    : num [1:18484] 0 4 5 5 3 0 4 0 1 2 ...
 $ Education        : chr [1:18484] "Bachelors" "Graduate Degree" "Bachelors" "High School"
...
 $ Occupation       : chr [1:18484] "Professional" "Management" "Management" "Professional"
...
 $ HomeOwnerFlag    : num [1:18484] 0 1 1 1 0 1 1 0 1 0 ...
 $ NumberCarsOwned  : num [1:18484] 1 2 3 3 0 1 4 3 4 2 ...

> summary(Clientes)
  TotalAmount        BikePurchase     CustomerID      Country
 Min.   :    2.29   Min.   :0.000   Min.   :11000   Length:18484
 1st Qu.:   49.97   1st Qu.:0.000   1st Qu.:15621   Class :character
 Median :  270.26   Median :0.000   Median :20242   Mode  :character
 Mean   : 1588.33   Mean   :0.494   Mean   :20242
 3rd Qu.: 2511.28   3rd Qu.:1.000   3rd Qu.:24862
 Max.   :13295.38   Max.   :1.000   Max.   :29483
 CountryRegionCode     Group              PersonID      PersonType
 Length:18484       Length:18484       Min.   : 1699   Length:18484
 Class :character   Class :character   1st Qu.: 6915   Class :character
 Mode  :character   Mode  :character   Median :11536   Mode  :character
                                       Mean   :11533
                                       3rd Qu.:16156
                                       Max.   :20777
 DateFirstPurchase               BirthDate                         Age
 Min.   :2001-07-01 00:00:00.00   Min.   :1910-08-13 00:00:00.000   Min.   : 40.00
 1st Qu.:2003-04-15 00:00:00.00   1st Qu.:1954-09-10 18:00:00.000   1st Qu.: 50.00
 Median :2003-11-05 00:00:00.00   Median :1963-08-14 00:00:00.000   Median : 57.00
 Mean   :2003-08-19 06:10:59.08   Mean   :1962-01-27 20:37:36.134   Mean   : 58.42
 3rd Qu.:2004-03-13 00:00:00.00   3rd Qu.:1970-09-26 00:00:00.000   3rd Qu.: 66.00
 Max.   :2004-07-31 00:00:00.00   Max.   :1980-12-26 00:00:00.000   Max.   :110.00
 MaritalStatus      YearlyIncome         Gender          TotalChildren
 Length:18484       Length:18484       Length:18484       Min.   :0.000
 Class :character   Class :character   Class :character   1st Qu.:0.000
 Mode  :character   Mode  :character   Mode  :character   Median :2.000
                                                          Mean   :1.844
                                                          3rd Qu.:3.000
                                                          Max.   :5.000
  Education          Occupation        HomeOwnerFlag    NumberCarsOwned
 Length:18484       Length:18484       Min.   :0.0000   Min.   :0.000
 Class :character   Class :character   1st Qu.:0.0000   1st Qu.:1.000
 Mode  :character   Mode  :character   Median :1.0000   Median :2.000
                                       Mean   :0.6764   Mean   :1.503
                                       3rd Qu.:1.0000   3rd Qu.:2.000
                                       Max.   :1.0000   Max.   :4.000
```

```
> colSums(is.na(Clientes))
      TotalAmount      BikePurchase        CustomerID           Country
                0                 0                 0                 0
CountryRegionCode             Group          PersonID        PersonType
                0                 0                 0                 0
DateFirstPurchase         BirthDate               Age     MaritalStatus
                0                 0                 0                 0
     YearlyIncome            Gender     TotalChildren         Education
                0                 0                 0                 0
       Occupation     HomeOwnerFlag   NumberCarsOwned
                0                 0                 0
```

Fuente: elaboración propia

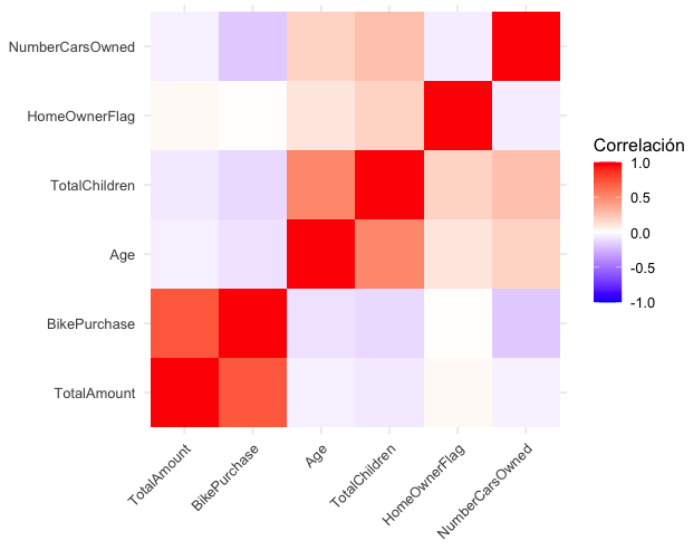
El "*dataset*" se compone de 18.484 registros y 19 variables; 8 numéricas, 9 categóricas y 2 en formato fecha. Además, observamos que no hay valores nulos.

Figura 2: Matriz de correlación

```
> #correlación
> variables_numericas  <- Clientes[sapply(Clientes, is.numeric)]
> correlaciones <- cor(variables_numericas, use = "complete.obs", method = "pearson")
> print(correlaciones)
                  TotalAmount BikePurchase    CustomerID      PersonID         Age
TotalAmount        1.00000000  0.724472470 -0.2476591513 -0.0018305200 -0.04927731
BikePurchase       0.72447247  1.000000000  0.0058039415 -0.0134880406 -0.09829074
CustomerID        -0.24765915  0.005803941  1.0000000000 -0.0002027785 -0.01361685
PersonID          -0.00183052 -0.013488041 -0.0002027785  1.0000000000 -0.00417936
Age               -0.04927731 -0.098290740 -0.0136168513 -0.0041793601  1.00000000
TotalChildren     -0.07380584 -0.127152246 -0.0060767714  0.0051763756  0.51644310
HomeOwnerFlag      0.02874607  0.007493770 -0.1247549970 -0.0085762664  0.10933082
NumberCarsOwned   -0.04474670 -0.180865163  0.0057038566  0.0132587326  0.18574240
                  TotalChildren HomeOwnerFlag NumberCarsOwned
TotalAmount       -0.073805844    0.028746071    -0.044746701
BikePurchase      -0.127152246    0.007493770    -0.180865163
CustomerID        -0.006076771   -0.124754997     0.005703857
PersonID           0.005176376   -0.008576266     0.013258733
Age                0.516443095    0.109330823     0.185742403
TotalChildren      1.000000000    0.185741534     0.265405072
HomeOwnerFlag      0.185741534    1.000000000    -0.057983482
NumberCarsOwned    0.265405072   -0.057983482     1.000000000
```

Fuente: elaboración propia

Figura 3: Matriz de correlación grafica

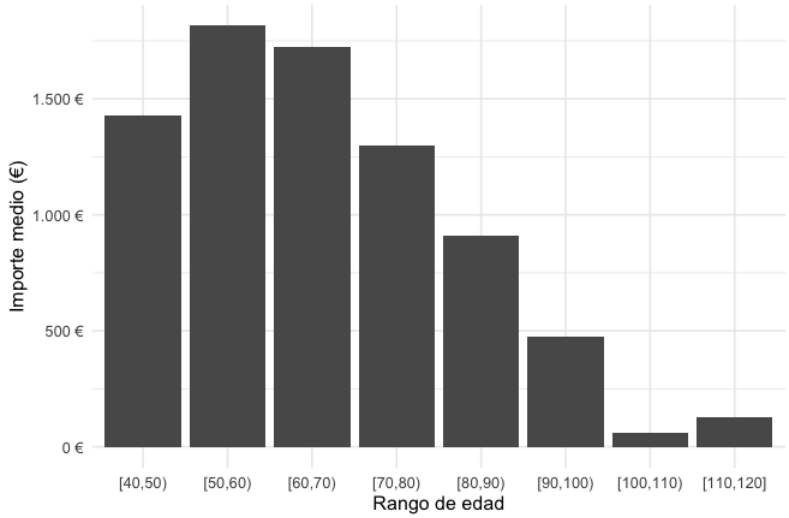


Fuente: elaboración propia

El análisis de correlación destaca dos relaciones positivas significativas:

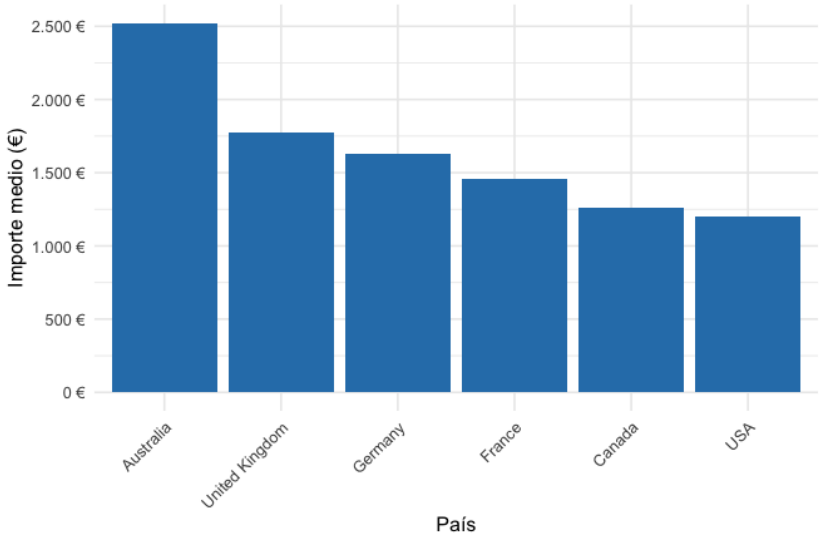
1. **Gasto y Compra de bicicleta:** Existe una relación directa entre el **gasto total** y la **compra de bicicletas**, lo que confirma que este artículo es el principal desembolso del cliente.
2. **Edad y Número de hijos:** Se observa una relación entre la **edad** y el **número de hijos**, vinculando a los perfiles de mayor edad con familias más amplias.

Figura 4: Importe medio por rango de edad



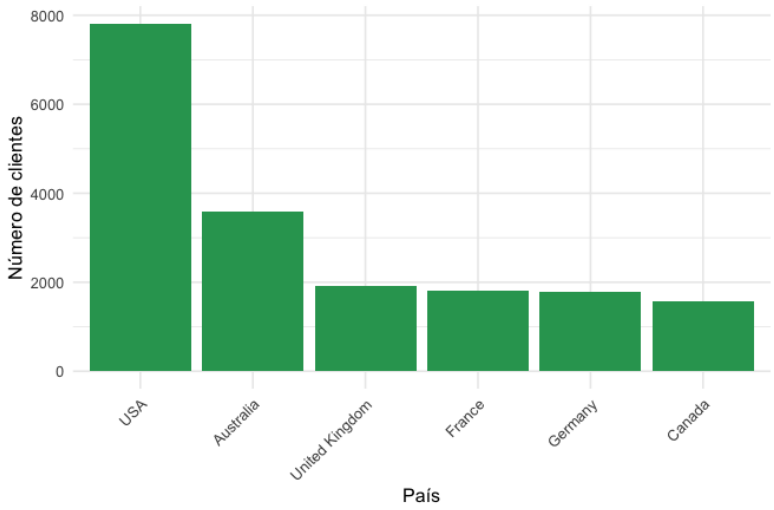
Fuente: elaboración propia

Figura 5: Importe medio por país



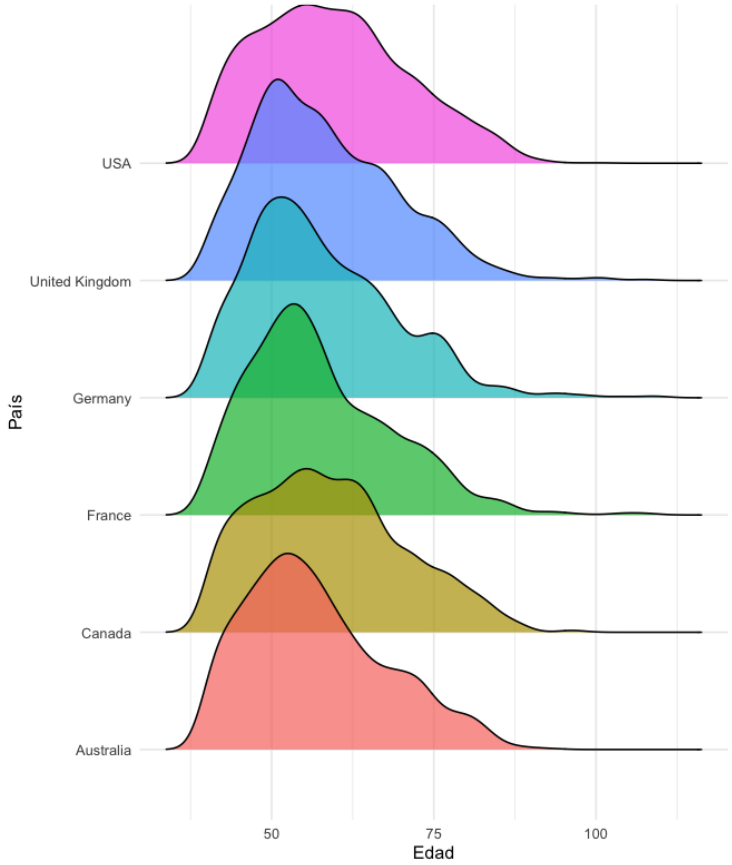
Fuente: elaboración propia

Figura 6: Número de clientes por país



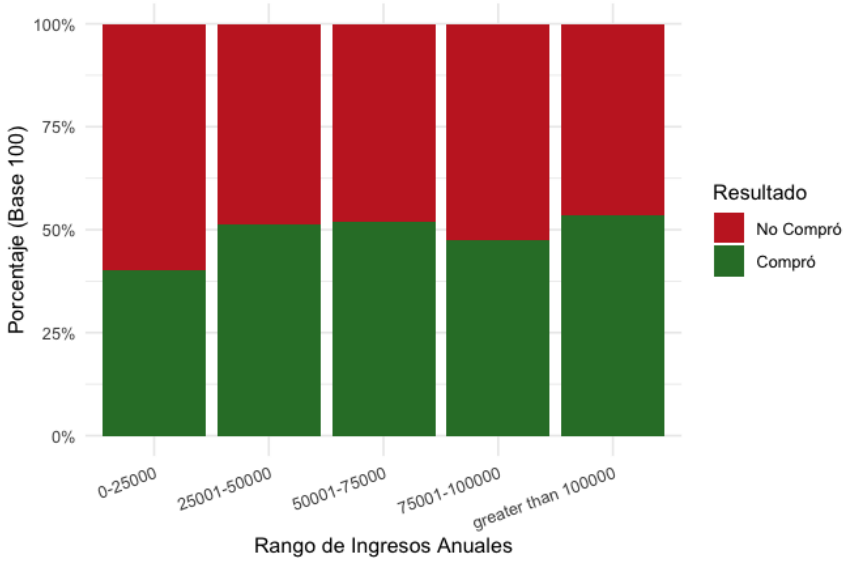
Fuente: elaboración propia

Figura 7: Distribución de clientes por edad y país



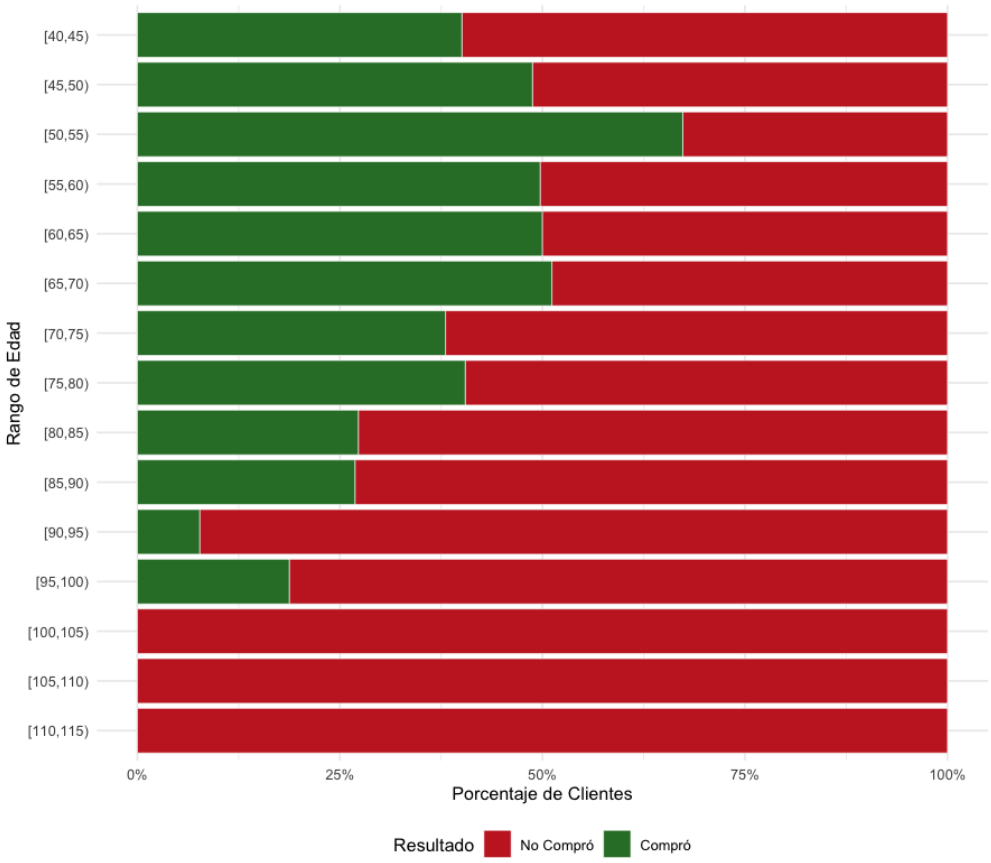
Fuente: elaboración propia

Figura 8: Proporción de compras por rango de ingresos



Fuente: elaboración propia

Figura 9: Proporción de compras por rango de edad



Fuente: elaboración propia

2. MODELOS DE APRENDIZAJE SUPERVISADO

2.1 Modelo de regresión logística

2.1.1 Modelo 1

Realizamos regresión logística con los datos de entrenamiento, previamente divididos en validación y entrenamiento:

Figura 10: Modelo de regresión logística 1

```
> regresionlogistica_ent1 <- glm(BikePurchase~
+                                TotalAmount+
+                                as.factor(Country)+
+                                Age+
+                                as.factor(Gender)+
+                                as.factor(MaritalStatus)+
+                                as.factor(YearlyIncome)+
+                                as.factor(TotalChildren)+
+                                as.factor(Education)+
+                                as.factor(Occupation)+
+                                as.factor(HomeOwnerFlag)+
+                                as.factor(NumberCarsOwned), data = entrenamiento, family="binomial")
```

```
Coefficients:
                                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                                -8.130e+00  1.446e+00  -5.623 1.88e-08 ***
TotalAmount                                 1.938e-02  7.146e-04  27.128  < 2e-16 ***
as.factor(Country)Canada                   -4.493e+00  5.551e-01  -8.095 5.73e-16 ***
as.factor(Country)Central                   1.378e+00  7.621e+00   0.181   0.8565
as.factor(Country)France                   -2.638e-01  7.832e-01  -0.337   0.7362
as.factor(Country)Germany                   5.699e-01  7.560e-01   0.754   0.4509
as.factor(Country)Northeast                -1.135e+01  1.206e+03  -0.009   0.9925
as.factor(Country)Northwest                 3.958e-01  5.005e-01   0.791   0.4291
as.factor(Country)Southeast                -1.009e+01  5.147e+02  -0.020   0.9844
as.factor(Country)Southwest                 7.601e-01  4.706e-01   1.615   0.1063
as.factor(Country)United Kingdom            4.041e-01  6.700e-01   0.603   0.5465
Age                                        -3.151e-03  1.777e-02  -0.177   0.8593
as.factor(Gender)M                         -1.029e-01  2.722e-01  -0.378   0.7054
as.factor(MaritalStatus)S                   4.327e-01  3.081e-01   1.404   0.1603
as.factor(YearlyIncome)25001-50000          5.432e-01  8.643e-01   0.629   0.5297
as.factor(YearlyIncome)50001-75000          9.186e-01  9.455e-01   0.972   0.3312
as.factor(YearlyIncome)75001-100000         5.554e-01  1.018e+00   0.546   0.5854
as.factor(YearlyIncome)greater than 100000  6.747e-01  1.218e+00   0.554   0.5795
as.factor(TotalChildren)1                   6.704e-01  4.958e-01   1.352   0.1764
as.factor(TotalChildren)2                   1.884e-01  5.044e-01   0.374   0.7088
as.factor(TotalChildren)3                   9.887e-01  5.224e-01   1.893   0.0584 .
as.factor(TotalChildren)4                   2.646e-01  5.424e-01   0.488   0.6257
as.factor(TotalChildren)5                   2.648e-01  7.647e-01   0.346   0.7291
as.factor(Education)Graduate Degree        -4.049e-01  4.827e-01  -0.839   0.4015
as.factor(Education)High School            -3.662e-01  4.846e-01  -0.756   0.4499
as.factor(Education)Partial College        -6.192e-01  4.324e-01  -1.432   0.1521
as.factor(Education)Partial High School    -1.131e-01  7.269e-01  -0.156   0.8763
as.factor(Occupation)Management            -2.069e-01  8.734e-01  -0.237   0.8127
as.factor(Occupation)Manual                 1.068e+00  9.221e-01   1.159   0.2466
as.factor(Occupation)Professional          -3.076e-01  6.911e-01  -0.445   0.6562
as.factor(Occupation)Skilled Manual         5.000e-01  5.671e-01   0.882   0.3780
as.factor(HomeOwnerFlag)1                  -5.299e-01  3.253e-01  -1.629   0.1033
as.factor(NumberCarsOwned)1                 1.530e-01  4.994e-01   0.306   0.7594
as.factor(NumberCarsOwned)2                -3.509e-01  5.452e-01  -0.644   0.5198
as.factor(NumberCarsOwned)3                -8.378e-01  8.327e-01  -1.006   0.3144
as.factor(NumberCarsOwned)4                -6.391e-01  9.209e-01  -0.694   0.4877
```

Fuente: elaboración propia

Tras analizar los coeficientes del modelo, se observa que la variable TotalAmount presenta una significancia estadística muy elevada. Este fenómeno se explica porque, al ser las bicicletas los artículos más caros de la

tienda, un monto total de gasto elevado determina en gran medida la variable dependiente (Y). En la práctica, el gasto actúa casi como un reflejo directo del hecho de haber comprado una bicicleta, lo que termina "opacando" el peso real de los demás factores.

Si bien su inclusión aumenta la precisión técnica, nuestro objetivo principal es profundizar en la comprensión del perfil del cliente y descubrir qué otros factores socioeconómicos influyen en la decisión de compra. Por este motivo, para evitar que el peso del gasto total esconda el efecto de variables como la educación, el país o la ocupación, optamos por excluir *TotalAmount* del análisis.

A continuación, se presenta un nuevo modelo de regresión generado sin esta variable ni tampoco las variables *Gender* y *HomeOwnerFlag* , con el fin de obtener una visión más equilibrada y detallada de los factores que caracterizan a nuestros compradores.

2.1.2 Modelo 2

Figura 11 Modelo de regresión logística 2

```
regresionlogistica_ent <- glm(BikePurchase~
                              as.factor(Country)+
                              Age+
                              as.factor(MaritalStatus)+
                              as.factor(YearlyIncome)+
                              as.factor(TotalChildren)+
                              as.factor(Education)+
                              as.factor(Occupation)+
                              as.factor(NumberCarsOwned), data = entrenamiento, family="binomial")
summary(regresionlogistica_ent)
```

```
Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                                1.134902   0.154317   7.354 1.92e-13 ***
as.factor(Country)Canada                  -0.887616   0.076176 -11.652  < 2e-16 ***
as.factor(Country)Central                 -1.453323   0.877875  -1.656 0.097823 .
as.factor(Country)France                  -0.676583   0.077721  -8.705  < 2e-16 ***
as.factor(Country)Germany                 -0.539019   0.079278  -6.799 1.05e-11 ***
as.factor(Country)Northeast               -1.109828   0.766760  -1.447 0.147778
as.factor(Country)Northwest               -0.920682   0.062000 -14.850  < 2e-16 ***
as.factor(Country)Southeast               -1.039676   0.693452  -1.499 0.133802
as.factor(Country)Southwest               -0.678586   0.058408 -11.618  < 2e-16 ***
as.factor(Country)United Kingdom          -0.462803   0.077605  -5.964 2.47e-09 ***
Age                                       -0.012513   0.002287  -5.471 4.47e-08 ***
as.factor(MaritalStatus)S                  0.301523   0.037186   8.109 5.12e-16 ***
as.factor(YearlyIncome)25001-50000         0.881553   0.095221   9.258  < 2e-16 ***
as.factor(YearlyIncome)50001-75000         1.310367   0.109398  11.978  < 2e-16 ***
as.factor(YearlyIncome)75001-100000        1.218424   0.118043  10.322  < 2e-16 ***
as.factor(YearlyIncome)greater than 100000 1.915581   0.141946  13.495  < 2e-16 ***
as.factor(TotalChildren)1                  0.318813   0.061707   5.167 2.38e-07 ***
as.factor(TotalChildren)2                  0.271344   0.063158   4.296 1.74e-05 ***
as.factor(TotalChildren)3                  0.249947   0.068500   3.649 0.000263 ***
as.factor(TotalChildren)4                 -0.261127   0.070568  -3.700 0.000215 ***
as.factor(TotalChildren)5                 -0.365898   0.085511  -4.279 1.88e-05 ***
as.factor(Education)Graduate Degree       -0.391280   0.059566  -6.569 5.07e-11 ***
as.factor(Education)High School           -0.099806   0.063569  -1.570 0.116403
as.factor(Education)Partial College       -0.145906   0.056038  -2.604 0.009223 **
as.factor(Education)Partial High School   -0.486567   0.091831  -5.298 1.17e-07 ***
as.factor(Occupation)Management           -0.288555   0.109459  -2.636 0.008384 **
as.factor(Occupation)Manual                0.121530   0.102879   1.181 0.237488
as.factor(Occupation)Professional         -0.332906   0.086245  -3.860 0.000113 ***
as.factor(Occupation)Skilled Manual       -0.341809   0.070970  -4.816 1.46e-06 ***
as.factor(NumberCarsOwned)1               -0.584312   0.061445  -9.510  < 2e-16 ***
as.factor(NumberCarsOwned)2               -0.901101   0.067324 -13.384  < 2e-16 ***
as.factor(NumberCarsOwned)3               -1.267098   0.090653 -13.977  < 2e-16 ***
as.factor(NumberCarsOwned)4               -1.881111   0.110032 -17.096  < 2e-16 ***
```

Fuente: elaboración propia

Figura 12 Cálculo del VIF (Variance Inflation Factor)

```
> vif(regresionlogistica_ent2)
                              GVIF Df GVIF^(1/(2*Df))
as.factor(Country)        2.698416  9        1.056697
Age                       2.136280  1        1.461602
as.factor(MaritalStatus)  1.133399  1        1.064612
as.factor(YearlyIncome)  20.970195  4        1.462852
as.factor(TotalChildren)  3.051508  5        1.118025
as.factor(Education)      3.977739  4        1.188378
as.factor(Occupation)    22.566301  4        1.476327
as.factor(NumberCarsOwned) 5.542764 4        1.238699
```

Fuente: elaboración propia

Se evaluó la independencia de las variables mediante el Factor de Inflación de la Varianza (VIF). Los resultados muestran que todos los valores de la métrica ajustada (GVIF^(1/(2*Df))) se encuentran por debajo de 1.5, situándose muy lejos del umbral crítico de 2.0. Esto confirma la ausencia de multicolinealidad en el modelo, garantizando que las variables predictoras no son redundantes y que los coeficientes estimados son estadísticamente fiables para el análisis del perfil del cliente.

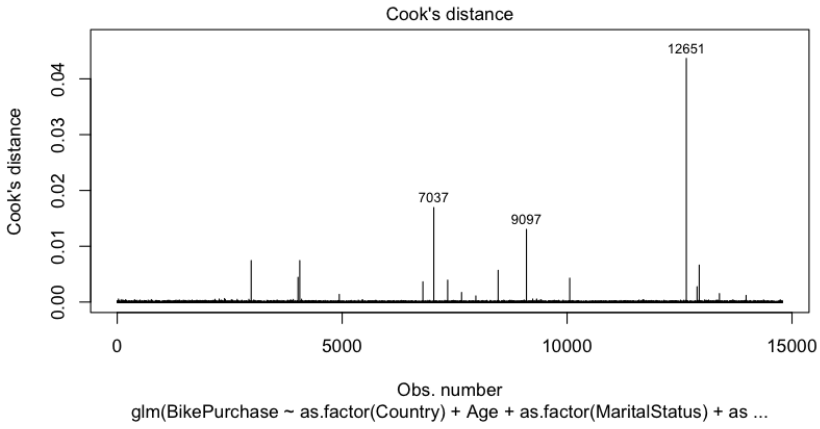
Figura 13: Test de Hosmer-Lemeshow

```
> hoslem.test(entrenamiento$BikePurchase, fitted(regresionlogistica_ent2))

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  entrenamiento$BikePurchase, fitted(regresionlogistica_ent2)
X-squared = 102.09, df = 8, p-value < 2.2e-16
```

Fuente: elaboración propia

El test de Hosmer-Lemeshow arrojó un p-valor < 0.05, lo que indica que el modelo no presenta un ajuste matemático perfecto entre las predicciones y los valores observados. Este resultado es habitual en modelos aplicados a grandes volúmenes de datos, donde el test se vuelve extremadamente sensible a pequeñas desviaciones. No obstante, dado que el objetivo del análisis es la identificación de perfiles y tendencias tras excluir la variable dominante *TotalAmount*, el modelo se considera plenamente válido y útil para la interpretación estratégica del comportamiento del cliente.

Figura 14: Distancia de Cook



Fuente: elaboración propia

Se calculó la Distancia de Cook para identificar posibles valores atípicos que pudieran sesgar los resultados del modelo. Como se observa en la gráfica, todos los registros presentan valores inferiores a 0.05, situándose muy por debajo del umbral crítico de 0.5. Esto confirma que el modelo no está influenciado por observaciones anómalas y que las tendencias identificadas en el perfil del cliente son estables y representativas de la muestra general.

2.1.2.1 Análisis de los coeficientes del modelo

Los resultados del modelo permiten trazar un perfil muy específico del comprador de bicicletas. Lo primero que salta a la vista es que el **nivel de ingresos** es el predictor de éxito más sólido; específicamente, aquellos clientes que superan los 100,000 anuales muestran la mayor disposición de compra, seguidos de cerca por el rango de ingresos medios.

Este interés se ve reforzado si el cliente es soltero, ya que este **estado civil** presenta una probabilidad de compra significativamente más alta que la de las personas casadas. En cuanto al entorno familiar, existe un "punto de equilibrio" interesante: tener entre uno y tres hijos incentiva la compra, pero al llegar a cuatro

o más, la tendencia se invierte drásticamente, probablemente por una cuestión de prioridades de gasto o logística familiar.

Por el contrario, el mayor freno para la venta de bicicletas es, sin duda, la **cantidad de vehículos** que ya posee el cliente en su hogar. Se observa una caída muy grande en la probabilidad de compra a medida que aumenta el número de coches, lo que sugiere que el automóvil funciona como un sustituto directo tanto en el transporte como en el ocio. A esto se suma que el interés decae conforme aumenta la **edad** del consumidor y en ciertas categorías profesionales o educativas más altas (como posgrados o puestos de gerencia), sectores que parecen optar por otras alternativas de recreación. **Geográficamente**, las regiones fuera del mercado principal muestran un rendimiento menor, lo que indica que el éxito de las ventas sigue muy vinculado a nichos geográficos y estilos de vida concretos.

Podríamos definir el cliente Ideal para la compra de bicicletas como:

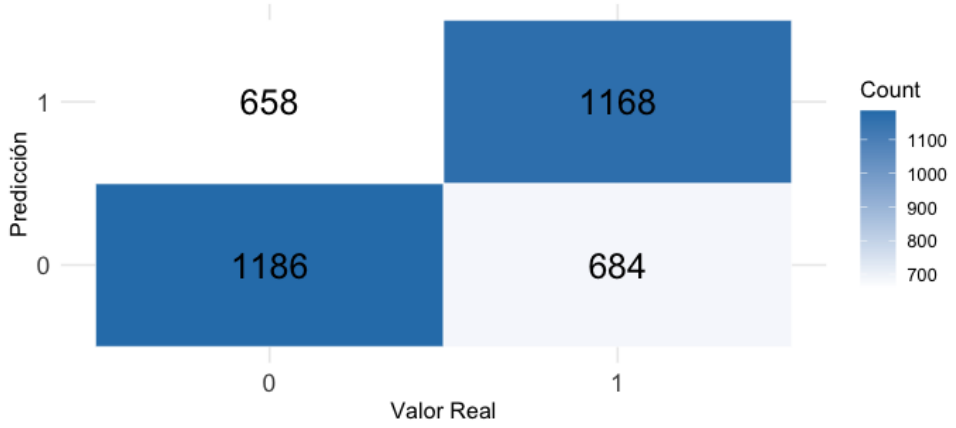
Un cliente joven, soltero, con ingresos altos (más de 100,000), que tiene entre 1 y 3 hijos y, fundamentalmente, que no posee autos (o tiene muy pocos).

Figura 15: Matriz de confusión (datos de validación)

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4780 2702
         1 2554 4752

               Accuracy : 0.6446
                 95% CI : (0.6368, 0.6523)
    No Information Rate : 0.5041
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.2892

 Mcnemar's Test P-Value : 0.0426

            Sensitivity : 0.6518
            Specificity : 0.6375
         Pos Pred Value : 0.6389
         Neg Pred Value : 0.6504
             Prevalence : 0.4959
         Detection Rate : 0.3232
   Detection Prevalence : 0.5060
      Balanced Accuracy : 0.6446

       'Positive' Class : 0
```

Fuente: elaboración propia

Figura 16: Matriz de confusión gráfica (datos de validación)



Fuente: elaboración propia

Se alcanza una **precisión global cercana al 64%**, un resultado razonable considerando que el análisis se basa exclusivamente en factores demográficos y socioeconómicos.

Se identifica simetría en los errores; el modelo falla casi con la misma frecuencia al predecir a un comprador que finalmente no lo es (658 casos) que al ignorar a un comprador real (684 casos).

En la práctica, esto significa que el modelo permite **identificar correctamente a 6 de cada 10 clientes potenciales**, proporcionando una base sólida para segmentar. Y a modo de ejemplo: lanzar campañas de marketing precisas y dirigidas sin necesidad de depender de datos de consumo previo.

Figura 17: Cálculo de *accuracy base*

```
> prop.table(table(entrenamiento$BikePurchase))

        0         1
0.5059508 0.4940492
```

Fuente: elaboración propia

El modelo alcanza una precisión del **63.7%**, superando el umbral de la 'apuesta tonta' (**50.59%**), lo que valida su utilidad al mejorar significativamente la capacidad de predicción sobre la simple tendencia de la clase mayoritaria.

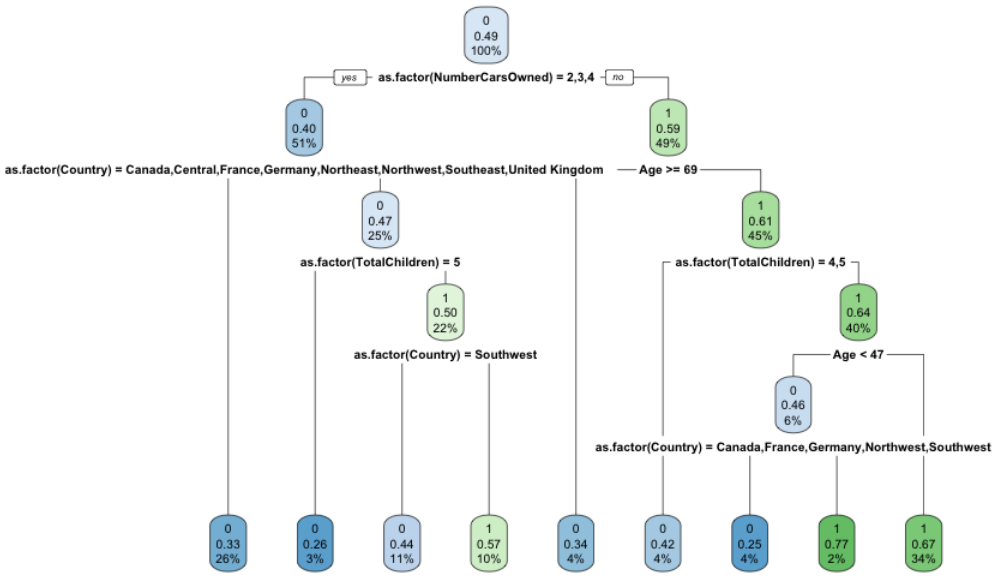
2.2 Modelo Árbol de Decisión

Figura 18: Modelo Árbol de decisión

```
modeloarbol <- rpart(as.factor(BikePurchase)~
                       as.factor(Country)+
                       Age+
                       as.factor(MaritalStatus)+
                       as.factor(YearlyIncome)+
                       as.factor(TotalChildren)+
                       as.factor(Education)+
                       as.factor(Occupation)+
                       as.factor(HomeOwnerFlag)+
                       as.factor(NumberCarsOwned),
                       data = entrenamiento, method = "class")
```

Fuente: elaboración propia

Figura 19: Modelo Árbol de Decisión (Gráfica)



Fuente: elaboración propia

Figura 20: Parámetros de complejidad

```
          CP nsplit rel error    xerror       xstd
1 0.17998905      0 1.0000000 1.0000000 0.008321743
2 0.02819600      1 0.8200109 0.8200109 0.008171138
3 0.01665298      2 0.7918149 0.7918149 0.008122896
4 0.01017429      5 0.7418560 0.7485628 0.008035339
5 0.01000000      8 0.7113332 0.7408979 0.008018078
```

Fuente: elaboración propia

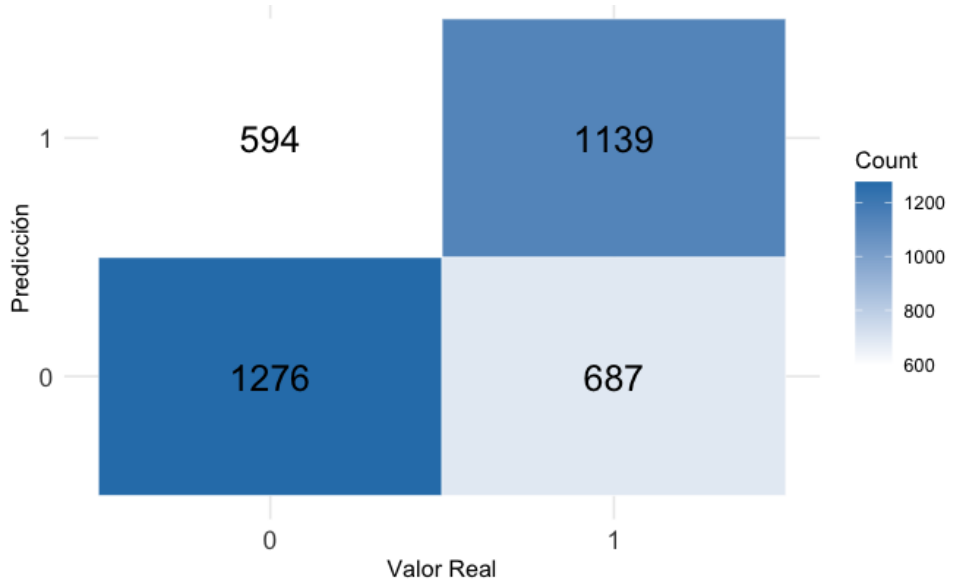
Se analizó la tabla de parámetros de complejidad (CP) para asegurar que el árbol no presente sobreajuste. Se observa que el error de validación cruzada (**xerror**) disminuye de forma constante a medida que aumenta el número de divisiones, alcanzando su punto más bajo (0.7408) con 8 divisiones. Esto confirma que el modelo actual captura patrones reales en los datos y mantiene una buena capacidad de generalización.

Figura 21: Matriz de Confusión (datos validación)

```
> confusionMatrix(as.factor(prediccion_validacion_cod),
+                       as.factor(validacion$BikePurchase))
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1276  687
         1  594 1139

               Accuracy : 0.6534
                 95% CI : (0.6378, 0.6688)
    No Information Rate : 0.506
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.3063

 Mcnemar's Test P-Value : 0.01016

            Sensitivity : 0.6824
            Specificity : 0.6238
         Pos Pred Value : 0.6500
         Neg Pred Value : 0.6572
             Prevalence : 0.5060
         Detection Rate : 0.3452
   Detection Prevalence : 0.5311
      Balanced Accuracy : 0.6531

       'Positive' Class : 0
```

Fuente: elaboración propia

Figura 22 Matriz de confusión grafica (datos validación)



Fuente: elaboración propia

El modelo alcanza una precisión del **65.34%**, clasificando correctamente a 2.415 individuos. Presenta un desempeño equilibrado, con una capacidad ligeramente superior para identificar con éxito a los no compradores (1,276 aciertos). Los niveles de error se mantienen estables en ambas categorías, lo que valida al árbol como una herramienta fiable para la segmentación estratégica de clientes, superando con claridad el rendimiento de la frecuencia base.

Figura 23: Importancia de las variables

Variable importance			
as.factor(NumberCarsOwned)	Age	as.factor(TotalChildren)	as.factor(Country)
23	17	17	16
as.factor(Education)	as.factor(YearlyIncome)	as.factor(Occupation)	
10	9	8	

Fuente: elaboración propia

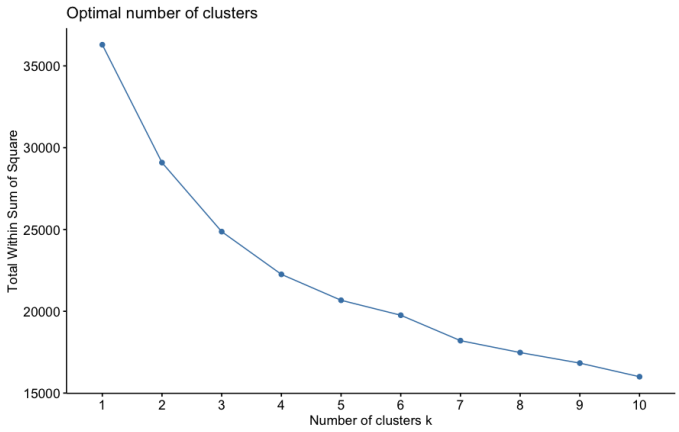
El modelo sitúa la cantidad de vehículos en propiedad (**NumberCarsOwned**) como el factor determinante con una puntuación de 23, estableciéndose como el principal criterio de clasificación. En un segundo nivel de relevancia se encuentran la edad (17), el número de hijos (17) y la ubicación geográfica (16), factores que presentan una influencia muy equilibrada entre sí. Por el contrario, el nivel educativo, los ingresos anuales y la ocupación muestran los pesos más bajos,

indicando que tienen un impacto significativamente menor en la estructura de decisión del modelo.

3. MODELOS DE APRENDIZAJE NO SUPERVISADO

3.1 Modelo de clusterización K-MEANS

Figura 24: Gráfica del Método del Codo



Fuente: elaboración propia

Se aplicó el método del codo para determinar el número óptimo de clústeres. Interpretamos que el número que nos ofrece una mayor cohesión interna y una mejor separación entre los segmentos es k=5.

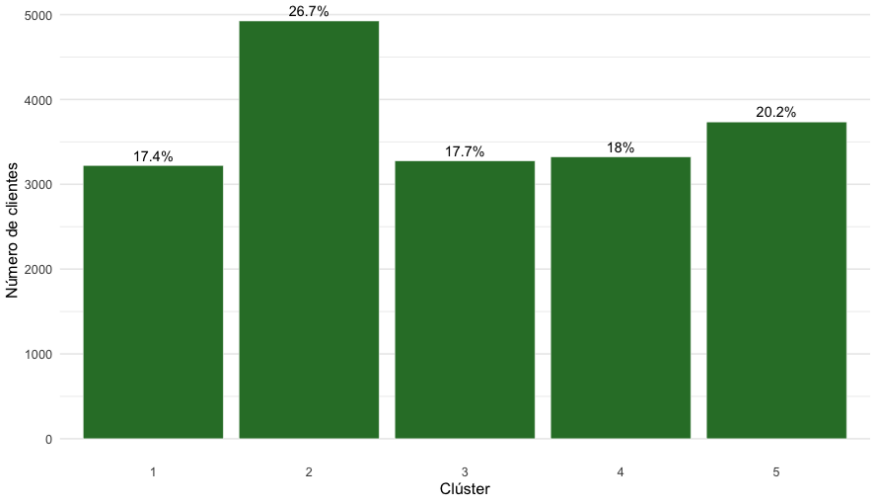
Figura 25: Matriz de medias estandarizadas de los clústeres

Cluster means:

	TotalAmount	BikePurchase	Age	TotalChildren	HomeOwnerFlag	NumberCarsOwned	Education_Num	Occupation_Num	YearlyIncome_Num
1	-0.5459482	-0.5952002	0.6385352	0.4724570	0.08871696	0.18710529	-0.89583744	-0.5840845	-0.7017828
2	0.2840854	0.6526129	-0.3718428	-0.3579727	0.39666296	-0.97353607	0.65913078	-0.3060638	-0.3307800
3	-0.4199607	-0.4077068	-1.0450819	-1.0157234	-0.81818549	-0.08480446	-0.64835821	-0.8525079	-0.7145446
4	1.3030903	1.0119465	0.2958307	0.1735888	-0.13647635	0.55126770	0.09922306	0.7727834	0.8801211
5	-0.6949452	-0.8901448	0.5932341	0.8012405	0.23963412	0.70636568	0.38403277	0.9677448	0.8853030

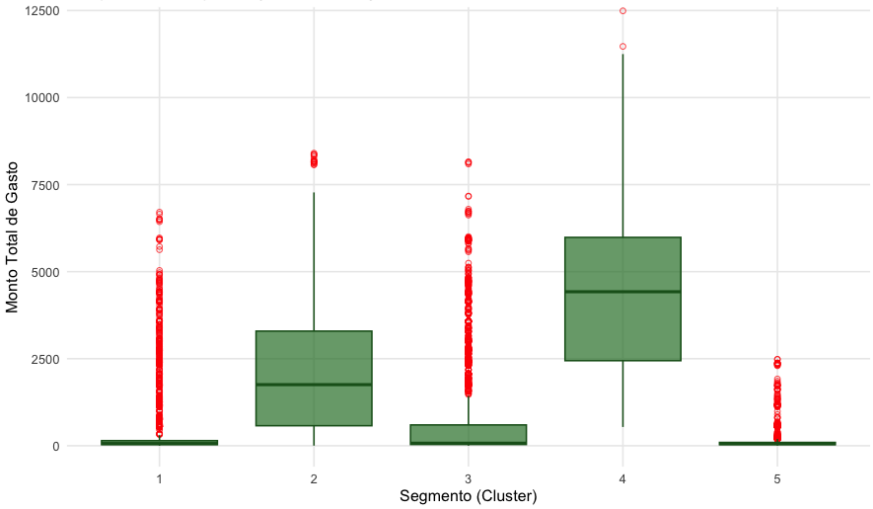
Fuente: elaboración propia

Figura 26: Número de clientes por Clúster



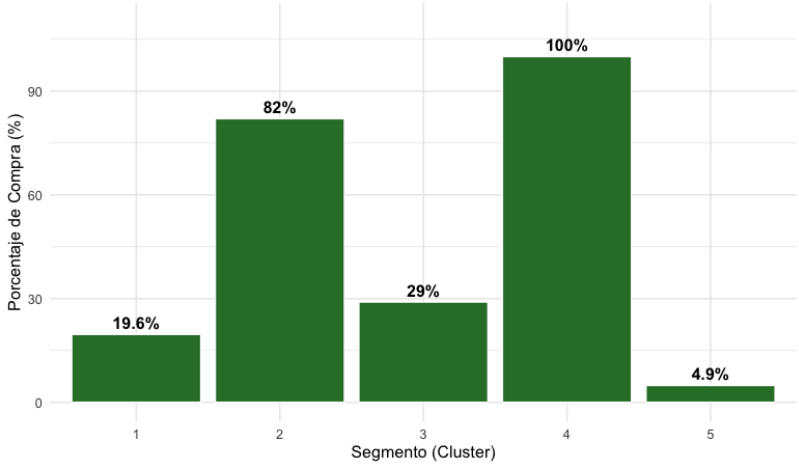
Fuente: elaboración propia

Figura 27: Distribución del TotalAmount por clúster



Fuente: elaboración propia

Figura 28: Tasa de compra de bicicleta por clúster



Fuente: elaboración propia

3.1.1 Interpretación de los grupos de clientes

Clúster 1: Perfil Maduro de Recursos Limitados (17,4%)

Este segmento agrupa a personas de mayor edad con niveles educativos y técnicos bajos. Aunque son propietarios de sus viviendas y tienen cargas familiares, sus ingresos son bastante reducidos (-0,70).

Clúster 2: El Motor de Ventas (26,7%)

Es el segmento más numeroso y el perfil estándar de éxito para el negocio. Se define como un grupo de profesionales urbanos con una alta compra de bicicletas (0,65). Su rasgo más distintivo es la mínima tenencia de vehículos (-0,97), lo que sugiere que utilizan la bicicleta como su principal medio de transporte o estilo de vida. Poseen niveles de educación y estatus profesional superiores a la media, consolidándose como el núcleo del volumen de ventas.

Clúster 3: Jóvenes en Crecimiento (17,7%)

Compuesto por los clientes más jóvenes del conjunto de datos (-1,04 en edad). Presentan ingresos limitados y no poseen vivienda propia, lo que se traduce en un gasto y un interés por la compra inferiores al promedio. Representan un mercado potencial a largo plazo, pero con una capacidad de conversión inmediata reducida debido a su etapa vital actual.

Clúster 4: El Segmento VIP (18,0%)

Representa la élite comercial de Adventure Works. Este grupo registra el gasto más elevado (1,30) y una probabilidad de compra absoluta (1,01). Son clientes de mediana edad con ingresos altos y perfiles profesionales cualificados. Aunque poseen vehículos, su alta capacidad económica les permite invertir en productos de gama alta, convirtiéndolos en el segmento más rentable por cliente.

Clúster 5: El Perfil de Alta Renta No Comprador (20,2%)

Este grupo es un hallazgo crítico: a pesar de tener los ingresos y puestos de trabajo más altos de la muestra (0,88 y 0,96), su disposición de compra es la más baja (-0,89). Su comportamiento está condicionado por un estilo de vida familiar maduro (muchos hijos) y una alta dependencia del automóvil (0,70), factores que actúan como barreras de entrada para la adquisición de bicicletas.

ÍNDICE DE FIGURAS