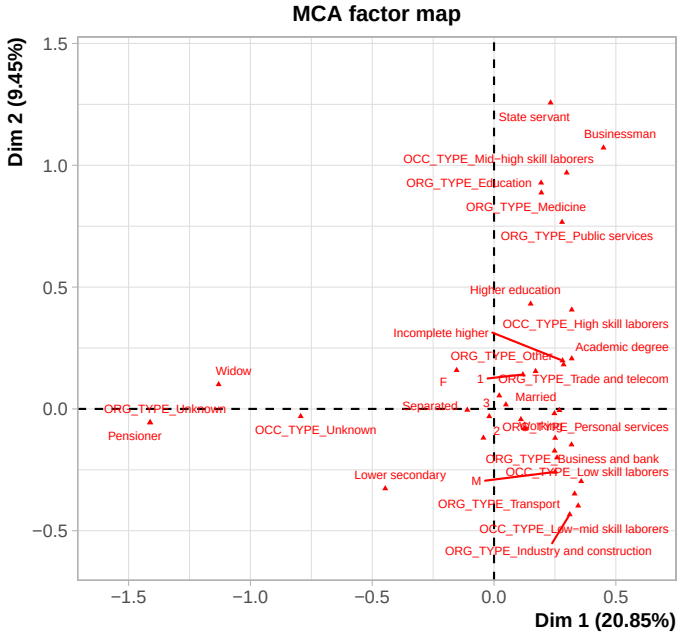
Análisis de correspondencias múltiples (ACM)

El Multiple Correspondance Analysis, ACM en adelante, es un método de análisis factorial para variables categóricas que permite analizar relaciones entre variables, así como reducir la dimensionalidad de la base de datos seleccionando sólo aquellas variables relevantes. Para realizarlo se deben escoger unas variables activas y otras de complementarias. En este caso, como el número de variables es relativamente bajo y se consideran todas relevantes, no se considerará ninguna variable complementaria. Además, se añadirán al análisis las variables numéricas como variables suplementarias, aunque solamente aquellas que se han considerado relevantes en el ACP.

Consideramos hacer una nueva codificación de las variables, reduciendo su longitud, de tal manera que los resultados obtenidos para el análisis se observen más claramente:

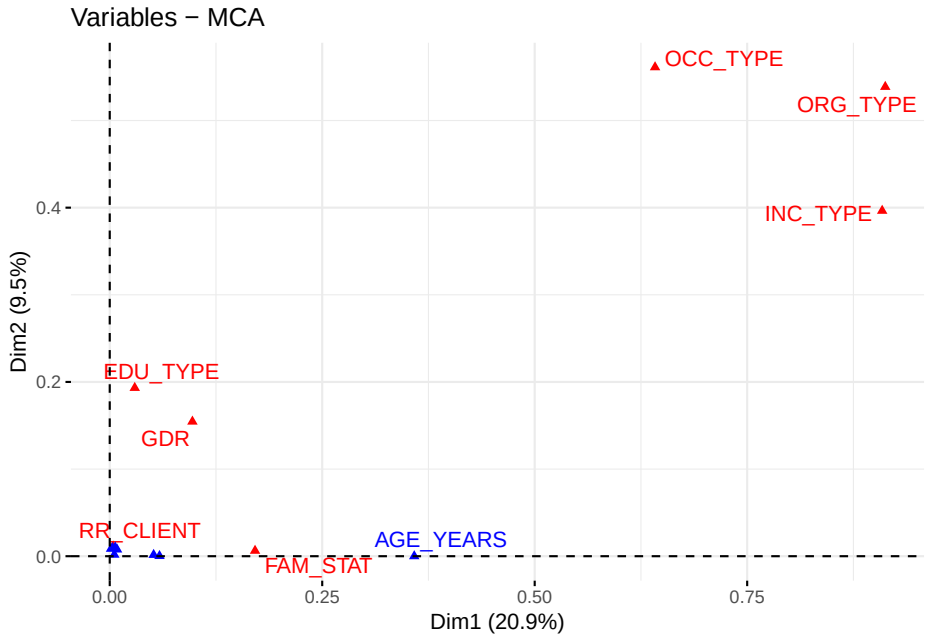
Desarrollo del ACM

Figura 63: Correlación entre Variables y Dimensiones Principales



En esta primera figura se representan las relaciones entre las modalidades de todas las variables categóricas con las dos primeras dimensiones del MCA. Se observa que la dimensión 1 se asocia con las variables que tienen relación con la edad, como la modalidad de Pensionista y Viudo. Se aprecia que la dimensión 2 se asocia a las modalidades relacionadas con la cualificación del trabajo del individuo, con una asociación positiva entre la dimensión y la cualificación del trabajador. Por tanto, se llamará a la dimensión 1 Edad, y a la dimensión 2 como cualificiación del trabajador.

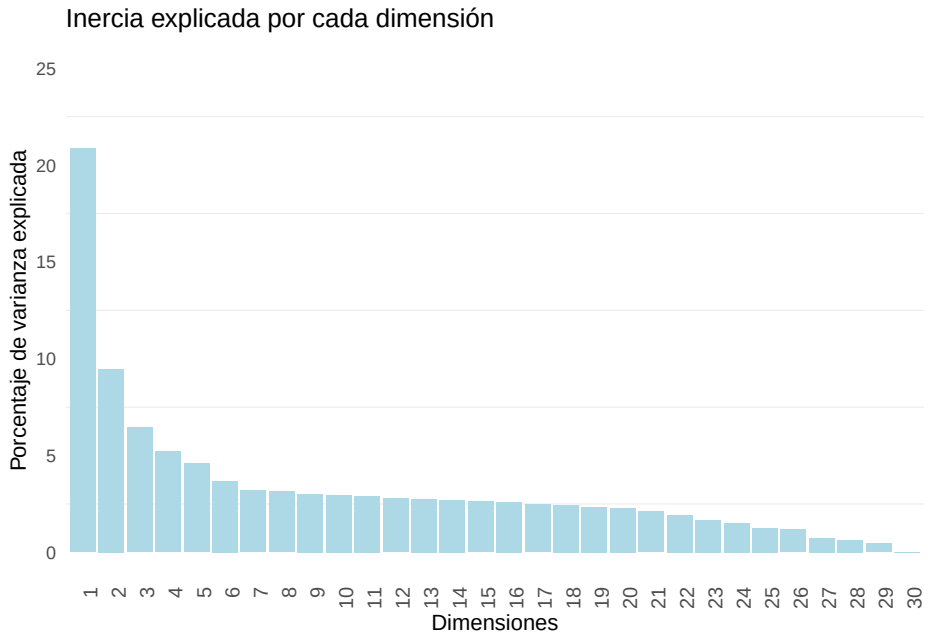
Figura 64: Variabilidad de las Variables Categóricas en las Dos Primeras Dimensiones



En el gráfico obtenido puede verse la variabilidad que expresan cada una de las variables categóricas en función de las dimensiones 1 y 2. Aquellas variables que estén más cerca del origen de coordenadas aportan muy poca información respecto a la variabilidad de los datos y, por tanto, son poco importantes. En cambio, aquellas variables más alejadas del centro aportan información más relevante.

Se representan gráficamente la inercia que explica cada una de las dimensiones generadas:

Figura 65: Inercia Explicada por cada Dimensión



Si una dimensión tiene una inercia baja, significa que todas las modalidades están muy cercanas al centro de gravedad y, en consecuencia, son muy similares. A medida que aumenta la inercia, va aumentando la distancia al centro de gravedad y, por tanto, se reduce la similitud.

Para poder estudiarlo más a fondo, se realiza la siguiente tabla en la que se puede observar para cada dimensión, su valor propio, el porcentaje de varianza (o inercia) explicada, y el porcentaje de varianza (o inercia) acumulada:

Cuadro 24: Varianza Explicada por cada Dimensión

	Valor propio	Porcentaje de la Varianza Acumulada	Porcentaje de Varianza
dim 1	0.16	20.85	20.85
dim 2	0.07	9.45	30.30
dim 3	0.05	6.45	36.76
dim 4	0.04	5.21	41.96
dim 5	0.03	4.59	46.56
dim 6	0.03	3.68	50.23
dim 7	0.02	3.19	53.42
dim 8	0.02	3.14	56.56
dim 9	0.02	3.00	59.57
dim 10	0.02	2.97	62.54
dim 11	0.02	2.91	65.45
dim 12	0.02	2.79	68.24
dim 13	0.02	2.74	70.98
dim 14	0.02	2.71	73.69
dim 15	0.02	2.63	76.32
dim 16	0.02	2.61	78.94
dim 17	0.02	2.48	81.41
dim 18	0.02	2.47	83.88
dim 19	0.02	2.33	86.21
dim 20	0.02	2.28	88.49
dim 21	0.02	2.12	90.61
dim 22	0.01	1.93	92.53
dim 23	0.01	1.68	94.21
dim 24	0.01	1.51	95.72
dim 25	0.01	1.25	96.97
dim 26	0.01	1.18	98.15
dim 27	0.01	0.74	98.90
dim 28	0.00	0.65	99.55
dim 29	0.00	0.45	100.00
dim 30	0.00	0.00	100.00

Tenemos un total de 31 dimensiones. La dimensión 1 destaca muy por encima del resto, explicando un 20.85 % de la variabilidad de los datos, seguida de la dimensión 2, explicando un 9.45 % de la variabilidad de los datos los datos. A partir de la dimensión 6, se ve que la gráfica se estabiliza bastante ya hasta la última dimensión.

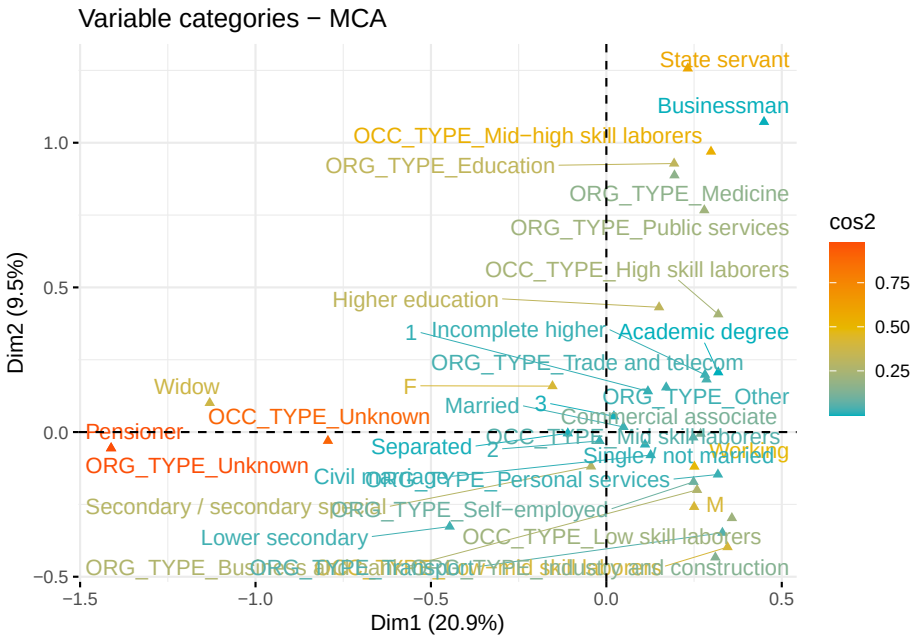
Por tanto, en total las dos primeras dimensiones ya explican un 30.3 % de la variabilidad de los datos, y se necesitan 17 dimensiones para llegar a tener una inercia acumulada por encima del 80 %.

Aunque las primeras dos dimensiones expliquen cerca del 30 % de la inercia, no todos los puntos se muestran igual de bien en las dos dimensiones. La calidad de la representación se llama coseno cuadrado (cos2), que mide el grado de asociación entre categorías de variables y un eje particular.

A continuación, se representa la calidad de las categorias a partir de ajustar los colores para cada punto proyectado, tomando como criterio el valor del coseno cuadrático (cos2). Si una categoría de variable está bien representada por dos dimensiones, la suma de cos2 es cercana a uno. Para algunos de los elementos de la fila, se requieren más de dos dimensiones para representar perfectamente los datos. Se considera lo siguiente:

- Las categorías de variables con valores bajos de cos2 se colorearán en "cian".
- Las categorías de variables con valores medios de cos2 se colorearán en "amarillo".
- Las categorías de variables con valores altos de cos2 se colorearán en "rojo".

Figura 66: Calidad de las Variables Categóricas a partir del Coseno Cuadrático

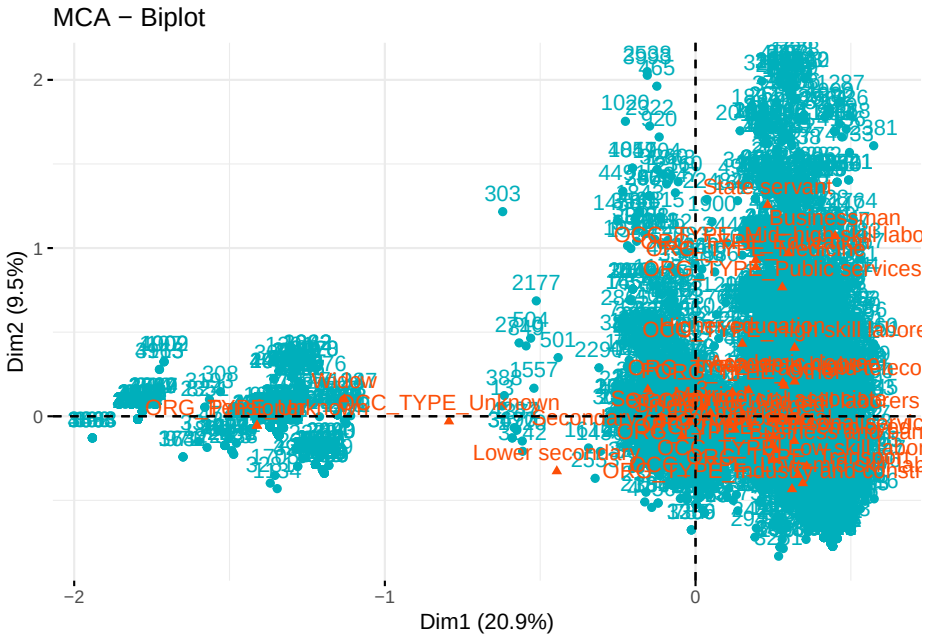


Salen muchas categorías que no están muy bien representadas por las dos primeras dimensiones. Esto implica que la posición de los puntos correspondientes en el diagrama de dispersión debe interpretarse con cierta cautela. Probablemente sea necesaria una solución de mayor dimensión. Aún así, se ha decidido no realizar el MCA de mayores dimensiones debido a la dificultad de representación gráfica. Por tanto, los resultados del MCA se analizarán con cautela, especialmente las variables poco representadas en las dos primeras dimensiones.

Gráfico de individuos y variables

Para una primera visualización de la relación entre las variables y las observaciones en un espacio reducido de dimensiones, acudimos al gráfico biplot. Este gráfico nos facilita la interpretación de la estructura de los datos al proporcionar una visualización que muestra la relación entre variables categóricas y observaciones.

Figura 67: Biplot de Individuos y Categorías

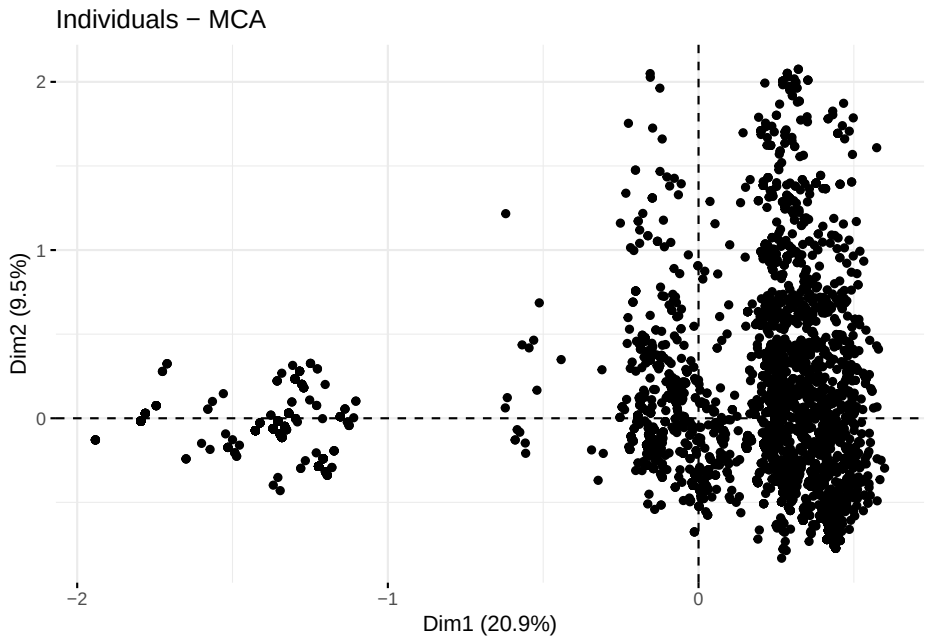


Con tal de poder analizar estos resultados de manera más precisa, se decide dividir este gráfico y analizar por una parte únicamente el gráfico de observaciones y por otra parte el gráfico de las variables categóricas.

Gráfico de individuos

Se representa gráficamente cómo se distribuyen los individuos en función de las dos primeras dimensiones que explican un 29.99 % de la variabilidad:

Figura 68: Gráfico de Individuos en las dos Primeras Dimensiones

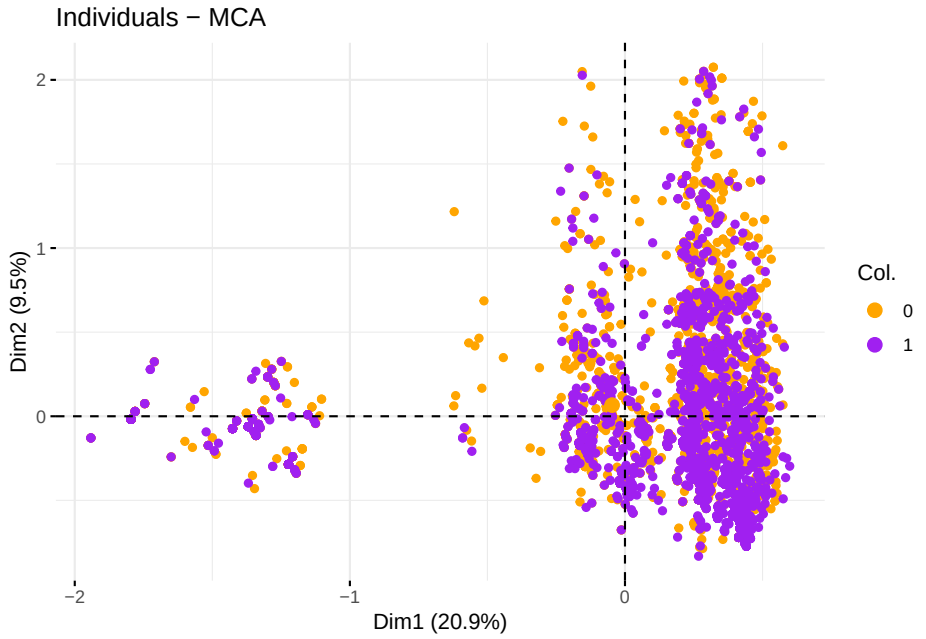


A simple vista, se aprecian varios grupos de individups pero resulta difícil distinguir cuántos. Sin embargo, sí podríamos decir que los individuos se dividen en como mínimo 2 grupos. Para distinguir mejor las agrupaciones de individuos y su asociación con algunas modalidades se pasa a estudiar cada variable para observar si existe algún tipo de asociación entre ellas.

Gráfico de los individuos según variable TARGET

A continuación representamos los mismos individuos pero coloreandolos según la variable "target", es decir, nuestra varible output, donde 1 indica aquel cliente con dificultades de pago, y 0 contrariamente:

Figura 69: Gráfico de Individuos según la Variable TARGET en las dos Primeras Dimensiones

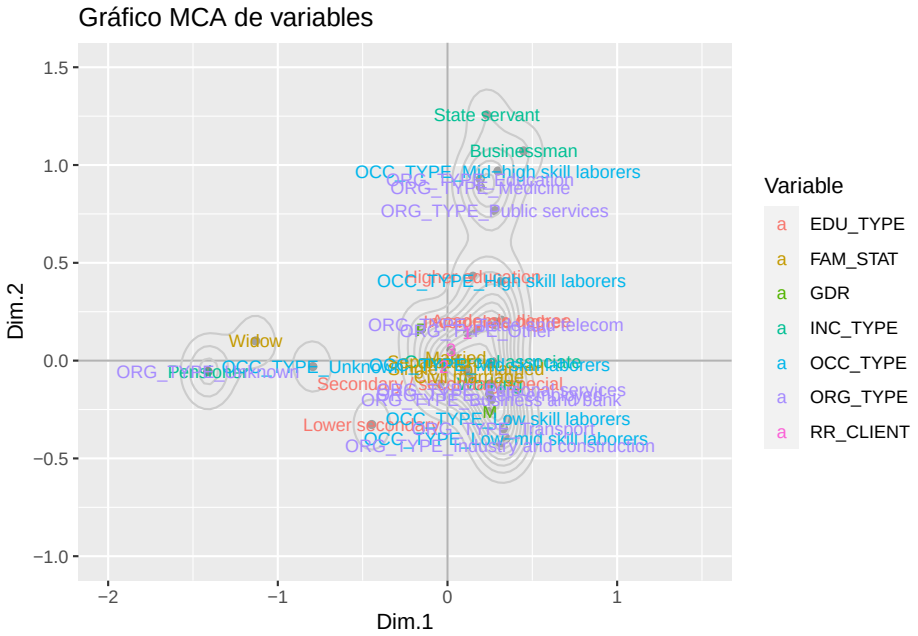


Observamos como diferenciando a los individuos según si tienen dificultades o no con el pago, no hay diferencias entre grupos de individuos, por lo tanto, podemos decir que no se ve ninguna asociación entre la variable TARGET y las modalidades de las variables representadas en estas dos dimensiones.

Gráfico de variables

Para tener una representación aún más clara sobre las variables y su asociación, a continuación se grafican estas variables con curvas de densidad para ver aquellas zonas donde hay una mayor concentración.

Figura 70: Representación de Variables en las Dimensiones del ACM



Como bien se ha comentado, la primera dimensión está asociada con las variables que tienen relación con la edad y la segunda dimensión se asocia a modalidades con la cualificación del trabajo del individuo.

Así pues, a partir de este gráfico de densidades, podemos ver como hay una relación muy destacada entre Widow y Pensioner en la dimensión 1. Esta correlación puede deberse a eventos de vida como la pérdida del cónyuge a una elevada edad y factores sociales, finalización laboral y por eso a una pensión por los años trabajados. La asociación de ambos términos con la población anciana y la edad es evidente.

En la dimensión 2 podemos ver una relación entre State Servant y Businessman con Mid/High Skill Laborers y con Education y Medicine. También podríamos ver relación con Higher education. La gente que trabaja en educación y/o mundo sanitario requieren un alto nivel de estudios y son trabajadores altamente cualificados. Al igual que podríamos asociarlo con los trabajadores en Servicios Públicos y Empresarios.

Aún y ver relación en este gráfico, hemos decidido realizar un gráfico de dispersión por cada una de las variables para ver si podíamos adquirir más información.

Gráficos de dispersión agrupado por cada variable

Con el objetivo de ver si las categorías son significativamente diferentes entre sí, se grafican gráficos de elipses alrededor de las categorías de cada una de las variables.

Se considerará que las categorías con elipses no superpuestas, es decir, separadas entre sí, son significativament diferentes entre sí. Por el contrario, cuando las elipses se superponen, nos indican que hay una similitud o asociación entre categorías, es decir, no són significativamente diferentes entre ellas.

Figura 71: Gráfico de Elipses NAME EDUCATION TYPE

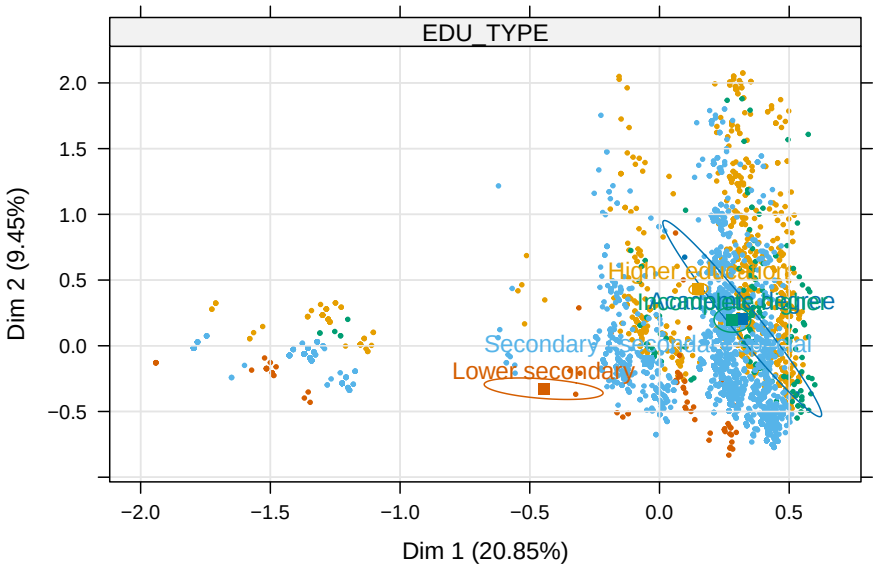


Figura 72: Gráfico de Elipses NAME FAMILY STATUS

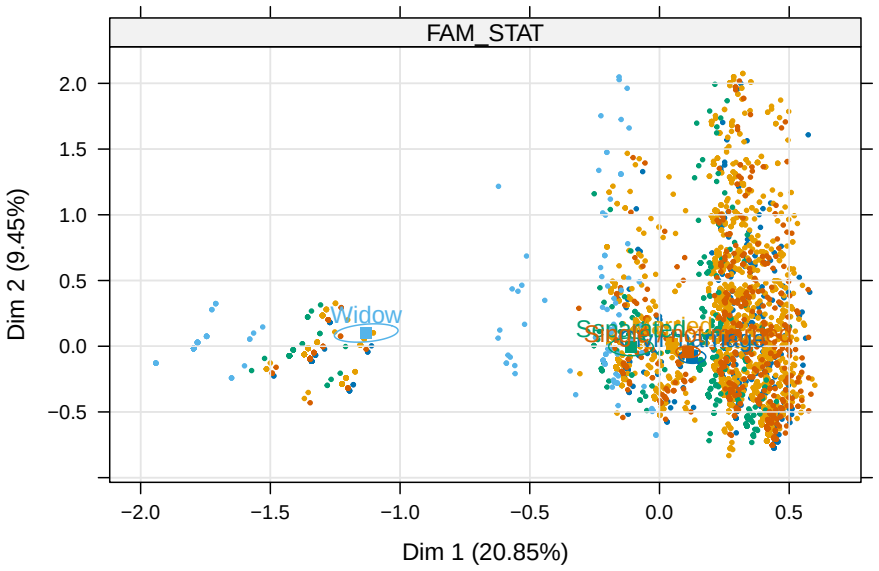


Figura 73: Gráfico de Elipses CODE GENDER

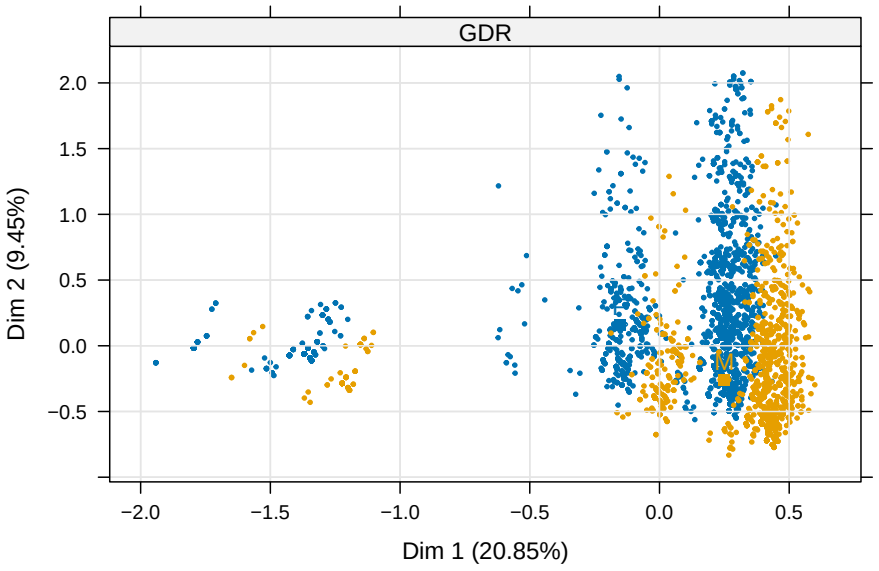


Figura 74: Gráfico de Elipses NAME INCOME TYPE

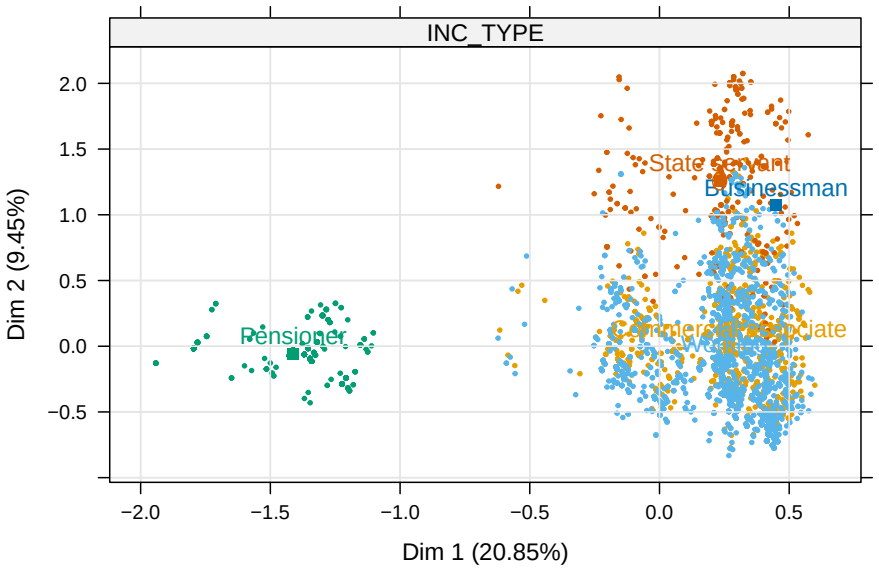


Figura 75: Gráfico de Elipses OCCUPATION TYPE

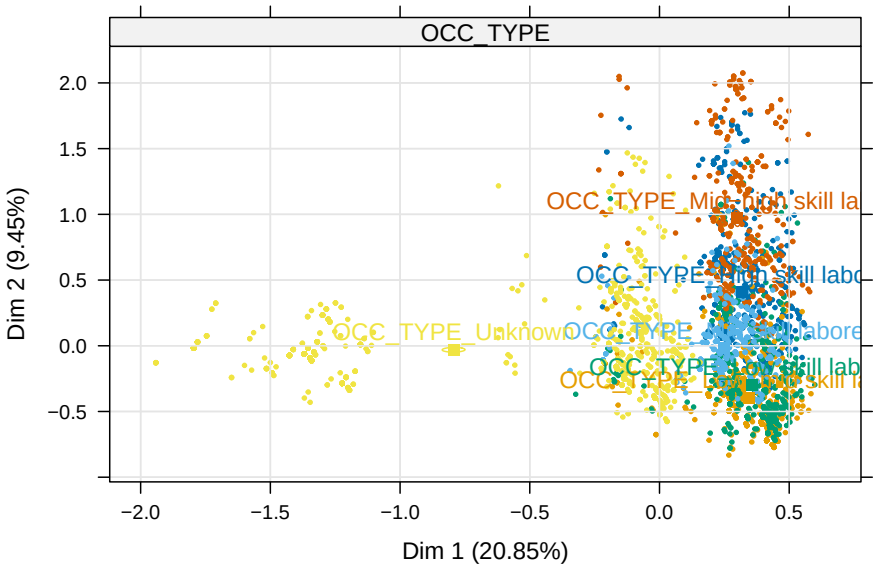


Figura 76: Gráfico de Elipses ORGANITATION TYPE

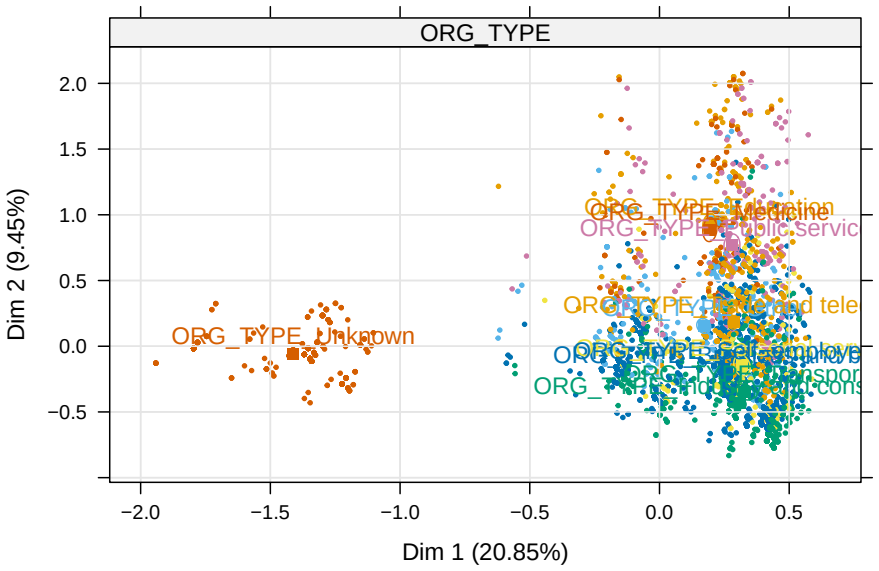
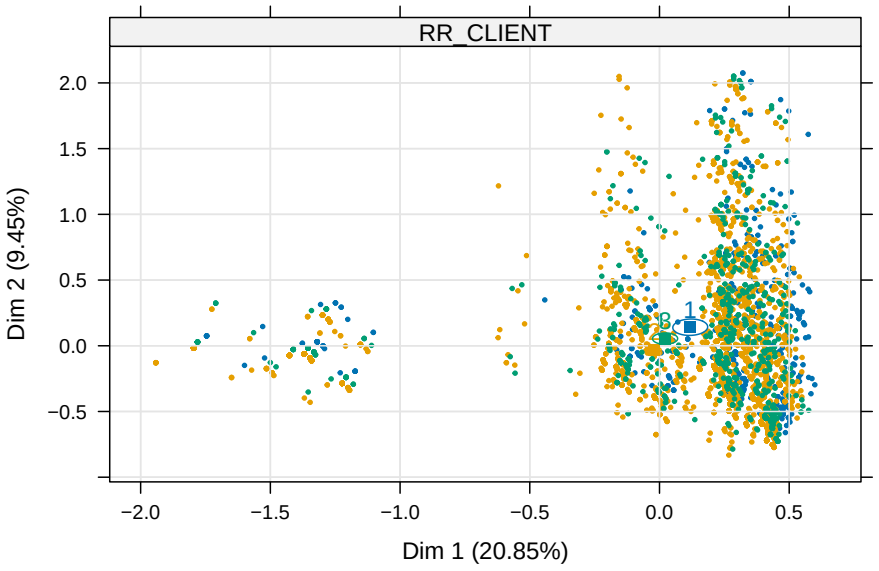


Figura 77: Gráfico de Elipses REGION RATING CLIENT



Al analizar cada variable de manerea individial en las dos primeras dimensiones del ACM, se observa que hay muchos gráficos donde encontramos categorías superpuestas entre ellas, y que por tanto, no nos aportan información significativa.

Como se ha observado que en las dos primeras dimensiones únicamente se muestra aproximadamente el 30 % de la variabilidad, para poder estudiar las categorías y su asociación más a fondo, hemos probado de analizar tres dimensions y no hi hay ningun gráfico significativo que nos permita extraer más información de la que ya hemos extraido con dos dimensiones. Por lo tanto, nos quedamos con los análisis sacados a partir de los gráficos de las dos primeras dimensiones.