ACP

2023-11-11

Table 1: Clase de cada variable

CODE_GENDER	factor
NAME_INCOME_TYPE	factor
NAME_EDUCATION_TYPE	factor
NAME_FAMILY_STATUS	factor
OCCUPATION_TYPE	factor
ORGANIZATION_TYPE	factor
REGION_RATING_CLIENT	factor
TARGET	factor
AMT_INCOME_TOTAL	numeric
AMT_CREDIT	numeric
AMT_ANNUITY	numeric
DAYS_BIRTH	numeric
OWN_CAR_AGE	numeric
AMT_GOODS_PRICE	numeric
CNT_FAM_MEMBERS	numeric
log_AMT_INCOME_TOTAL	numeric
log_AMT_CREDIT	numeric
log_AMT_ANNUITY	numeric
log_AMT_GOODS_PRICE	numeric
AGE_YEARS	numeric
DIFF_CREDIT_GOODS	numeric
RATIO_CREDIT_INCOME	numeric
RATIO_ANNUITY_CREDIT	numeric
DTI_RATIO	numeric

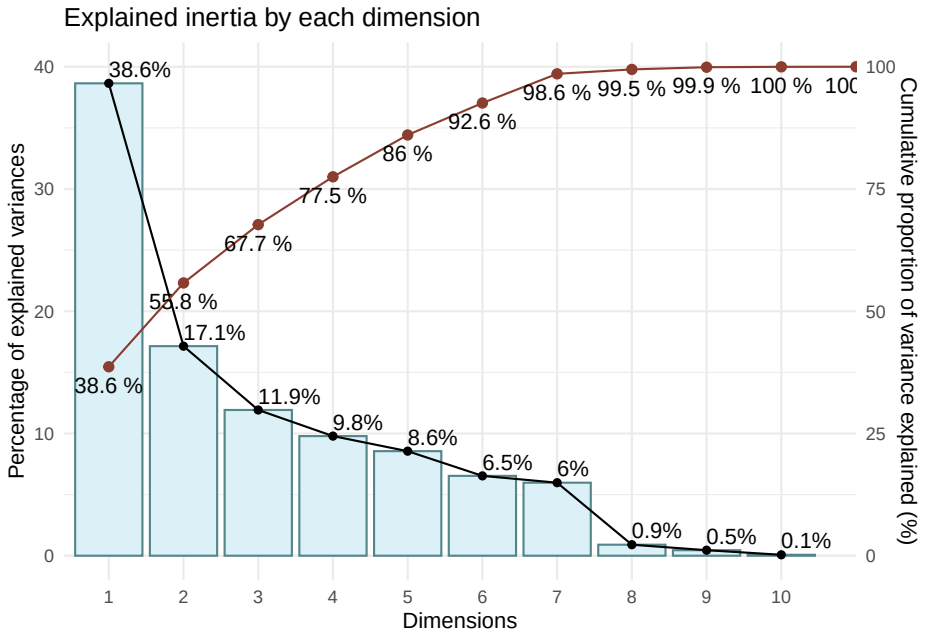
Se observa que la base de datos tiene un total de 11 columnas numéricas. Por tanto, el análisis de componentes principales tendrá como máximo 11 componentes.

Selección de variables numéricas

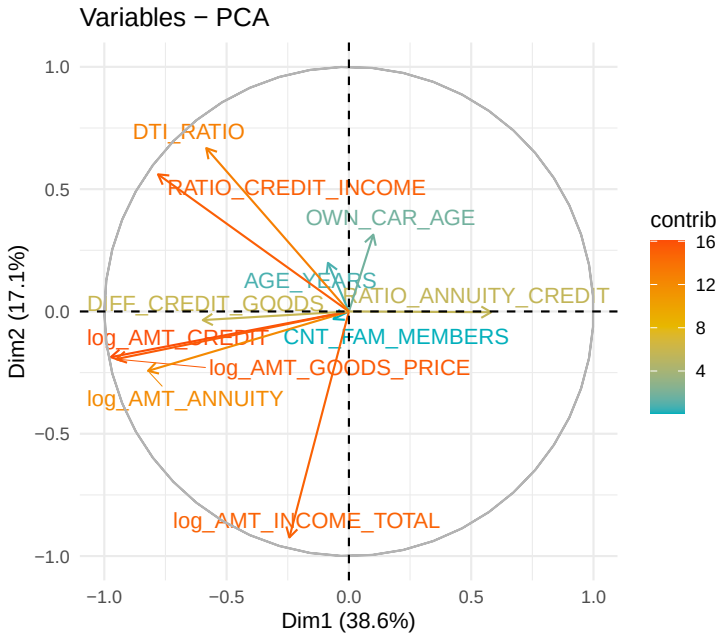
Se proceden a eliminar, primeramente, aquellas variables para las cuales ya existe su transformación logarítmica. Esto se hace para no contar con variables que contengan la misma capacidad explicativa (y así evitar colinealidad). También se elimina la variable DAYS_BIRTH, ya que se cuenta con AGE_YEARS, que es una transformación de la inicial, debido a que DAYS_BIRTH no tenia una clara interpretación.

PCA

A partir de aquí, se procede con el análisis de componentes principales.



Teniendo en cuenta que la inercia equivale a la proporción de la variabilidad de los datos, se sabe que con un 80% de inercia se puede obtener casi toda la información o variabilidad de la base de datos original. Con ello, vemos que el 80% de la inercia acumulada se logra con 5 planos factoriales, pero aún se pueden eliminar algunas variables.



Observamos la tabla de rotaciones:

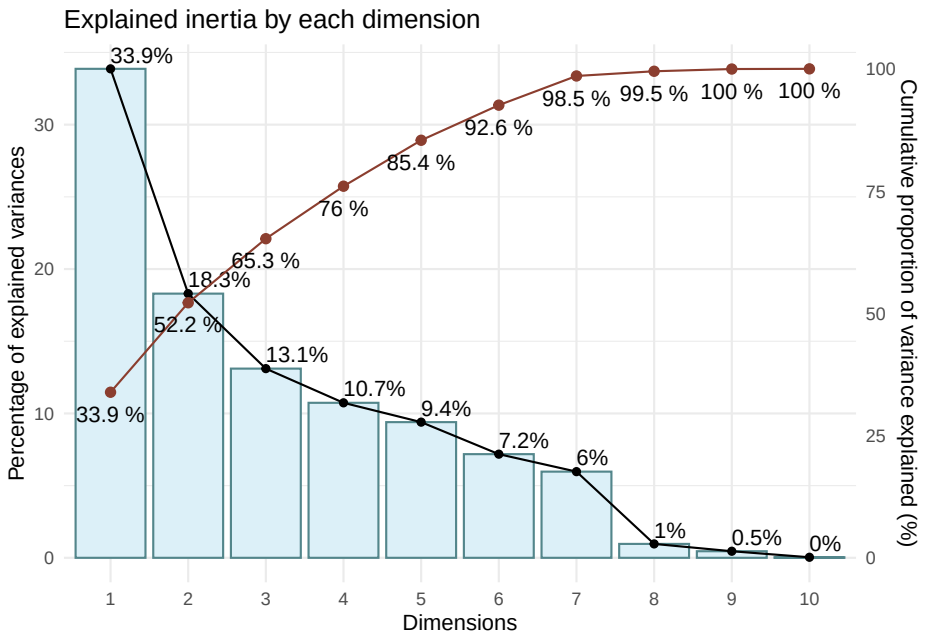
Table 2: Correlación de cada variable con cada plano factorial

	PC1	PC2	PC3	PC4	PC5
OWN_CAR_AGE	0.0483554	0.2287002	0.0530876	0.2413126	0.9271662
CNT_FAM_MEMBERS	-0.0305664	-0.0226378	0.6166566	0.3768438	-0.0869280
log_AMT_INCOME_TOTAL	-0.1180318	-0.6727750	-0.0175937	-0.0797217	0.1936175
log_AMT_CREDIT	-0.4721472	-0.1358924	-0.0277607	0.0061186	0.0394832
log_AMT_ANNUITY	-0.3983901	-0.1769402	0.2080437	-0.3959441	0.1684720
log_AMT_GOODS_PRICE	-0.4612147	-0.1418139	-0.0257791	-0.0460653	0.0317867
AGE_YEARS	-0.0415368	0.1447106	-0.6281559	-0.2150941	0.1247852
DIFF_CREDIT_GOODS	-0.2897238	-0.0256553	-0.0363394	0.3324768	0.0667128
RATIO_CREDIT_INCOME	-0.3784918	0.4081350	-0.0056133	0.0574366	-0.1162079
RATIO_ANNUITY_CREDIT	0.2819794	-0.0021710	0.3500729	-0.6037860	0.1736897
DTI_RATIO	-0.2828528	0.4863060	0.2310985	-0.3313722	-0.0259258

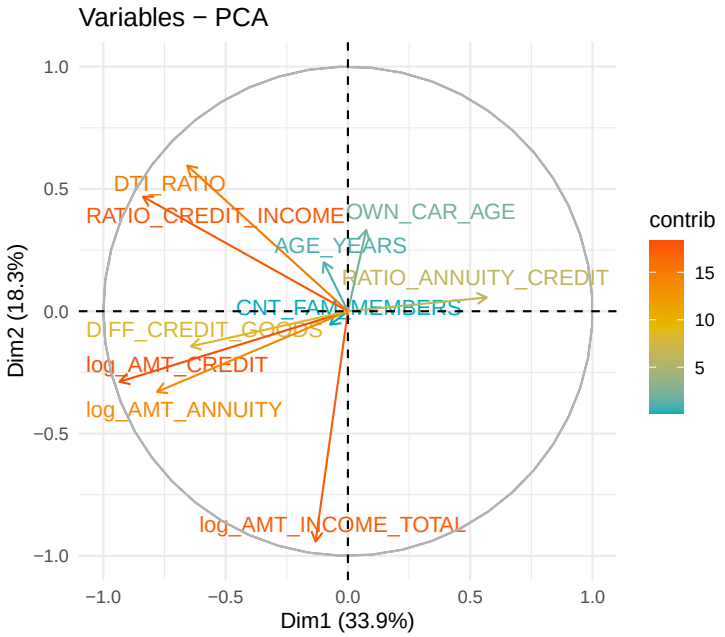
En el grafico vemos que las flechas de **log_AMT_GOODS_PRICE** y **log_AMT_CREDIT** se solapan entre ellas, eso queire decir que las dos variables explican el mismo plano factorial. Vemos en la tabla de rotaciones que **log_AMT_CREDIT** contribuye más a explicar el primer plano factorial, y además las correlaciones entra cada una de las variables y cada dimensión son muy similares. Por esta razón eliminamos **log_AMT_GOODS_PRICE**.

Nos quedamos con una variable menos, por tanto tenemos 10 variables numéricas.

De vuelta, verificamos el porcentaje de inercia por cada componente principal y la acumulada:



Como se puede ver, seguimos teniendo 5 dimensiones que acumulan el 80% de la varianza.



Vemos que las variables **CNT_FAM_MEMBERS**, **AGE_YEARS** y **OWN CAR AGE** no explican las dos primeras componentes pero si nos fijamos en la tabla de rotaciones vemos que sí tienen importancia a la hora de explicar las otras tres dimensiones:

Table 3: Correlación de cada variable con cada plano factorial

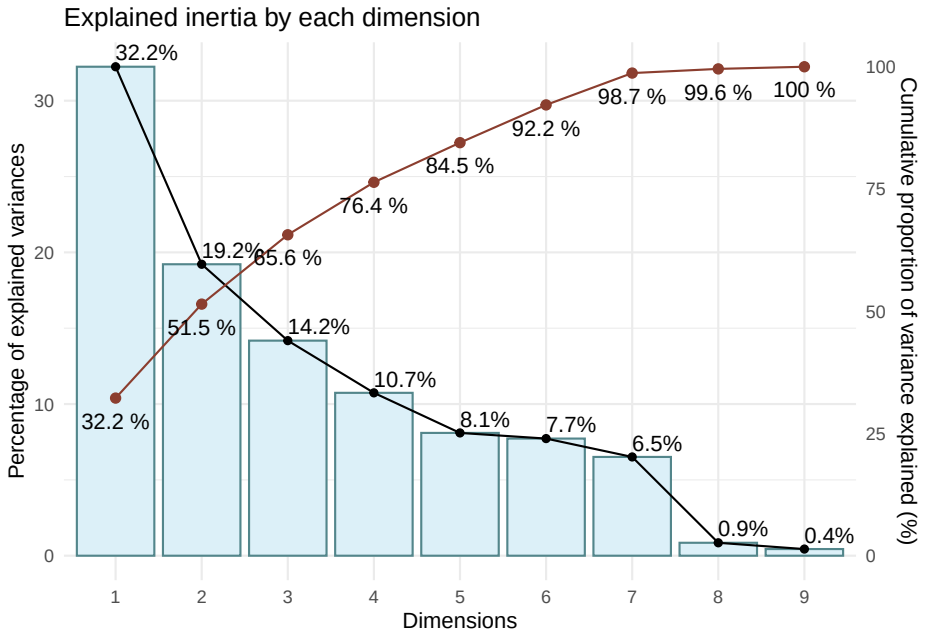
	PC1	PC2	PC3	PC4	PC5
OWN_CAR_AGE	0.0406536	0.2454602	0.0602964	0.2636258	0.9169085
CNT_FAM_MEMBERS	-0.0394947	-0.0404725	0.6131245	0.3812205	-0.0952529
log_AMT_INCOME_TOTAL	-0.0717491	-0.6961680	-0.0308119	-0.0950202	0.2049254
log_AMT_CREDIT	-0.5073525	-0.2130431	-0.0401256	-0.0102255	0.0487471
log_AMT_ANNUITY	-0.4240602	-0.2436766	0.1986781	-0.4109116	0.1850255
AGE_YEARS	-0.0537890	0.1481145	-0.6256961	-0.2195745	0.1310781
DIFF_CREDIT_GOODS	-0.3487325	-0.1061162	-0.0544017	0.3010831	0.0894349
RATIO_CREDIT_INCOME	-0.4552798	0.3462113	-0.0081212	0.0515503	-0.1143317
RATIO_ANNUITY_CREDIT	0.3084733	0.0416441	0.3591344	-0.5960684	0.1815479
DTI_RATIO	-0.3570194	0.4403417	0.2343324	-0.3322223	-0.0197281

Por ejemplo, en el caso de **OWN CAR AGE** se puede ver en la tabla anterior que, se podría decir que no es la que mejor explica las primeras componentes, pero vemos que explica casi toda la componente 5.

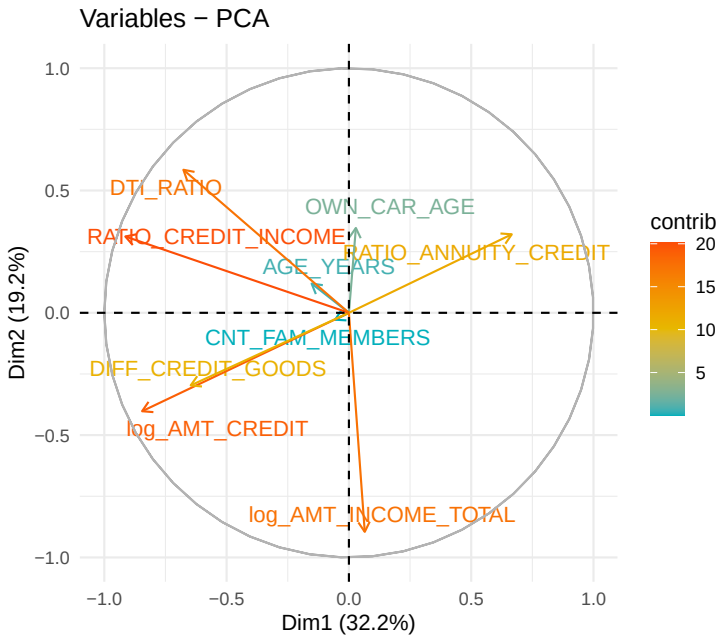
Otra observación se podria hacer de las variables **log_AMT_CREDIT** y **log_AMT_ANNUITY**, donde se puede apreciar que tienen correlaciones similares con la primera y segunda dimensión. Teniendo en cuenta que esas dos primeras dimensiones (PC1 y PC2) són las más importantes, ya que acumulan la mayoría de la inercia (en total un 52.2%), parece una decisión sensata eliminar una de ellas, en este caso **log_AMT_ANNUITY**.

Ahora conservamos 9 variables numéricas.

De forma igual que anteriormente, comprobamos el porcentaje de inercia para cada componente principal y la acumulada:



Como se puede comprobar, las 5 dimensiones siguen siendo las necesarias para acumular el 80% de la varianza.



Observamos tambien la tabla de rotaciones para verificar si se puede eliminar alguna variable más:

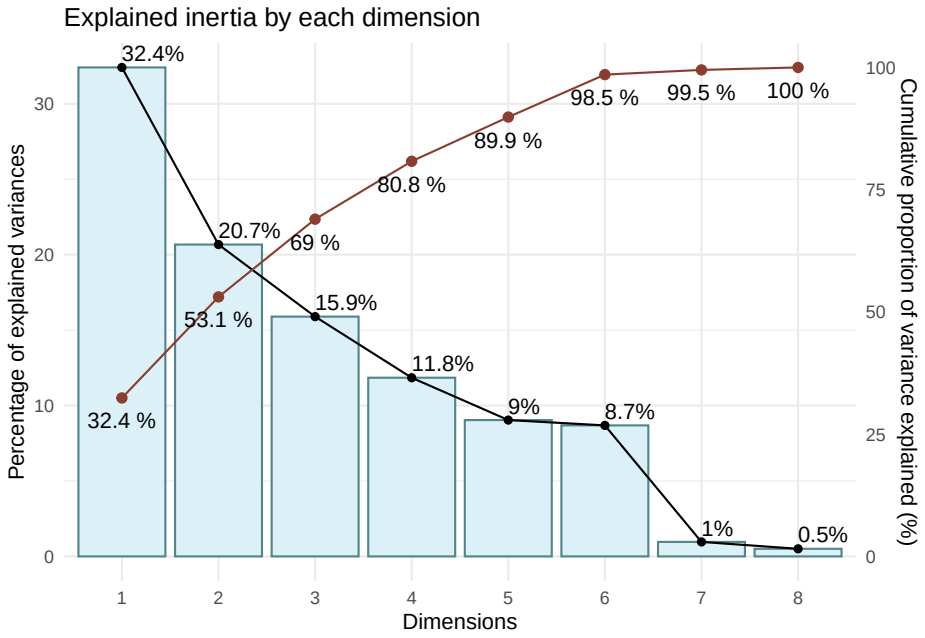
Table 4: Correlación de cada variable con cada plano factorial

	PC1	PC2	PC3	PC4	PC5
OWN_CAR_AGE	0.0167177	0.2638115	0.0414438	-0.9282134	-0.1613777
CNT_FAM_MEMBERS	-0.0330431	-0.0213960	0.7023098	-0.0655018	0.6542655
log_AMT_INCOME_TOTAL	0.0380475	-0.6808910	0.0089027	-0.0712942	-0.1829002
log_AMT_CREDIT	-0.4963663	-0.3058538	0.0179660	0.0098952	-0.1328703
AGE_YEARS	-0.0895071	0.0909410	-0.6812394	-0.0661836	0.5211075
DIFF_CREDIT_GOODS	-0.3797467	-0.2256788	0.0707076	-0.1959997	-0.1718557
RATIO_CREDIT_INCOME	-0.5366794	0.2372406	0.0189005	0.0813545	0.0015314
RATIO_ANNUITY_CREDIT	0.3908576	0.2440248	0.1423681	0.1802348	-0.3724869
DTI_RATIO	-0.3972254	0.4446957	0.1221834	0.2169086	-0.2344155

Si nos fijamos en el gráfico que incluye los dos primeros planos factoriales (PC1 y PC2), resulta fácil ver que **log_AMT_CREDIT** y **DIFF_CREDIT_GOODS** se solapan en su proyección, teniendo **log_AMT_CREDIT** más contribución dado que el vector es más largo. De aquí se entiende que las correlaciones de ambas variables en los dos primeros planos factoriales son muy similares, motivo por el cual solapan. En la tabla de correlaciones anterior se puede comprobar como efectivamente, estas correlaciones son similares. Incluso la correlación en ambas variables con la tercera dimensión (PC3) es baja, de forma parecida. Por tanto, se procede a eliminar aquella con menos contribución en PC1 y PC2, esta siendo **DIFF_CREDIT_GOODS**.

Ahora se conservan 8 variables numéricas.

Se vuelven a ejecutar todos los pasos anteriores para volver a verificar si hace falta eliminar más variables:



Se aprecia como la eliminación de **DIFF_CREDIT_GOODS** ha modificado el número de dimensiones necesarias para alcanzar el 80% de inercia acumulada, pasando de 5 a 4 dimensiones.

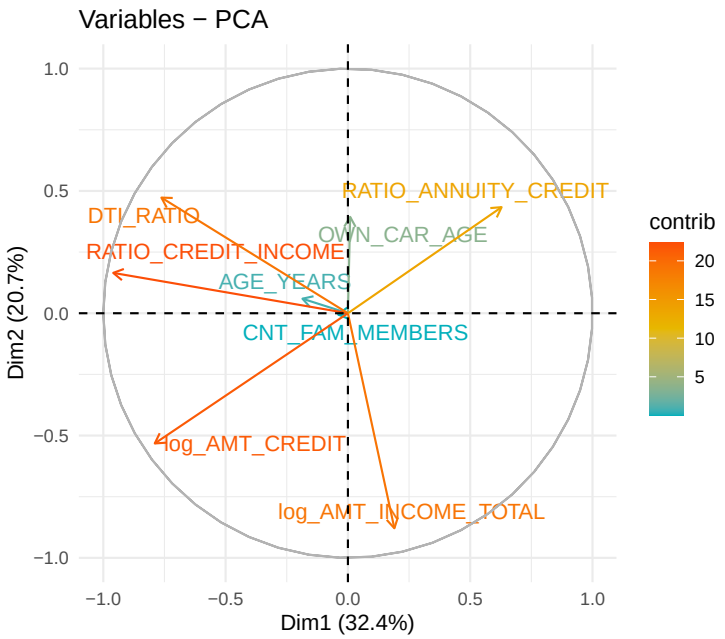


Table 5: Correlación de cada variable con cada plano factorial

	PC1	PC2	PC3	PC4
OWN_CAR_AGE	0.0056870	0.3062623	-0.0030588	0.9203547
CNT_FAM_MEMBERS	-0.0247561	-0.0041456	-0.7074073	0.1177313
log_AMT_INCOME_TOTAL	0.1179538	-0.6836391	-0.0420849	0.1218891
log_AMT_CREDIT	-0.4904066	-0.4139790	-0.0598471	0.0681678
AGE_YEARS	-0.1154462	0.0462637	0.6823184	0.0564692
RATIO_CREDIT_INCOME	-0.5958824	0.1282325	-0.0355111	-0.0417412
RATIO_ANNUITY_CREDIT	0.3902375	0.3372441	-0.1098849	-0.2629810
DTI_RATIO	-0.4735547	0.3675972	-0.1237687	-0.2132899

Comprobando el gráfico de las dos primeras dimensiones, y analizando las correlaciones, parece ser que ya no hace falta eliminar más variables. Por tanto, conservamos 8 variables numéricas.

Las variables eliminadas han sido: - **AMT_INCOME_TOTAL**, **AMT_CREDIT**, **AMT_ANNUITY**, **AMT_GOODS_PRICE**, todas ellas con motivo de que ya se habia creado otra variable a partir de su transformación logarítmica. - **DAYS_BIRTH**, ya que la variable **AGE_YEARS** es una transformación de ella. - **log_AMT_GOODS_PRICE** - **log_AMT_ANNUITY** - **DIFF_CREDIT_GOODS**

Interpretación de planos factoriales

Para ayudar a dar nombre a las diferentes dimensiones, aparte de utilizar las herramientas gráficas, también podemos fijarnos en las correlaciones entre las variables y los componentes principales.

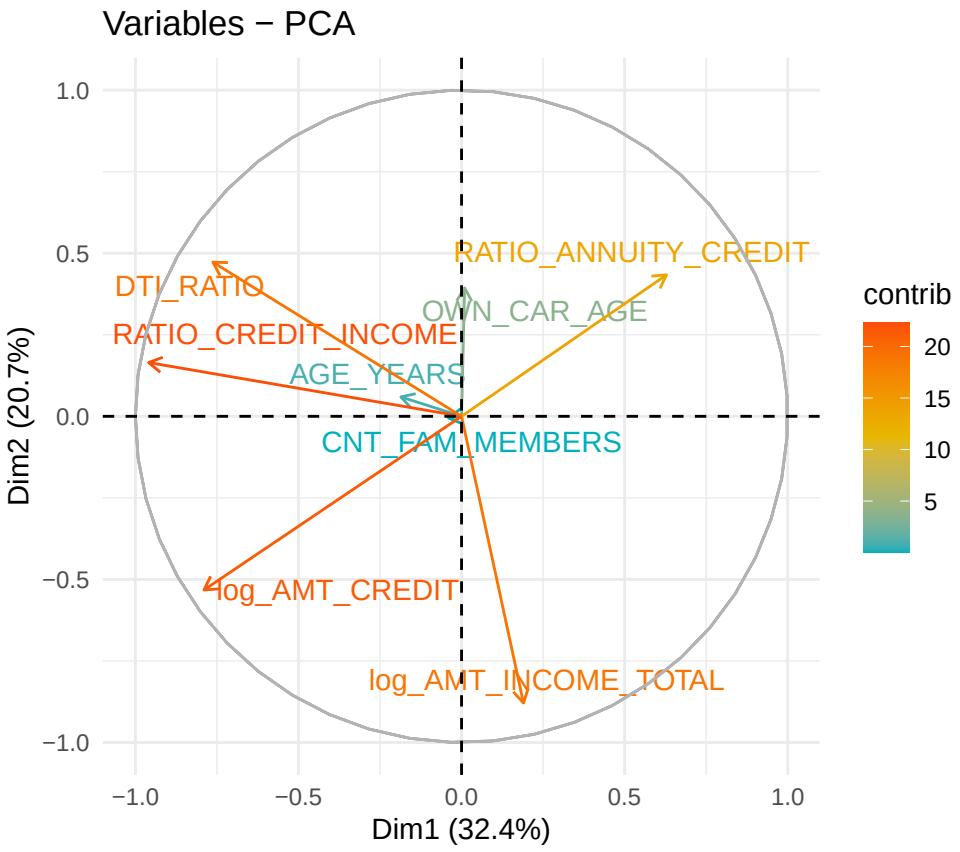


Table 6: Correlación de cada variable con cada plano factorial

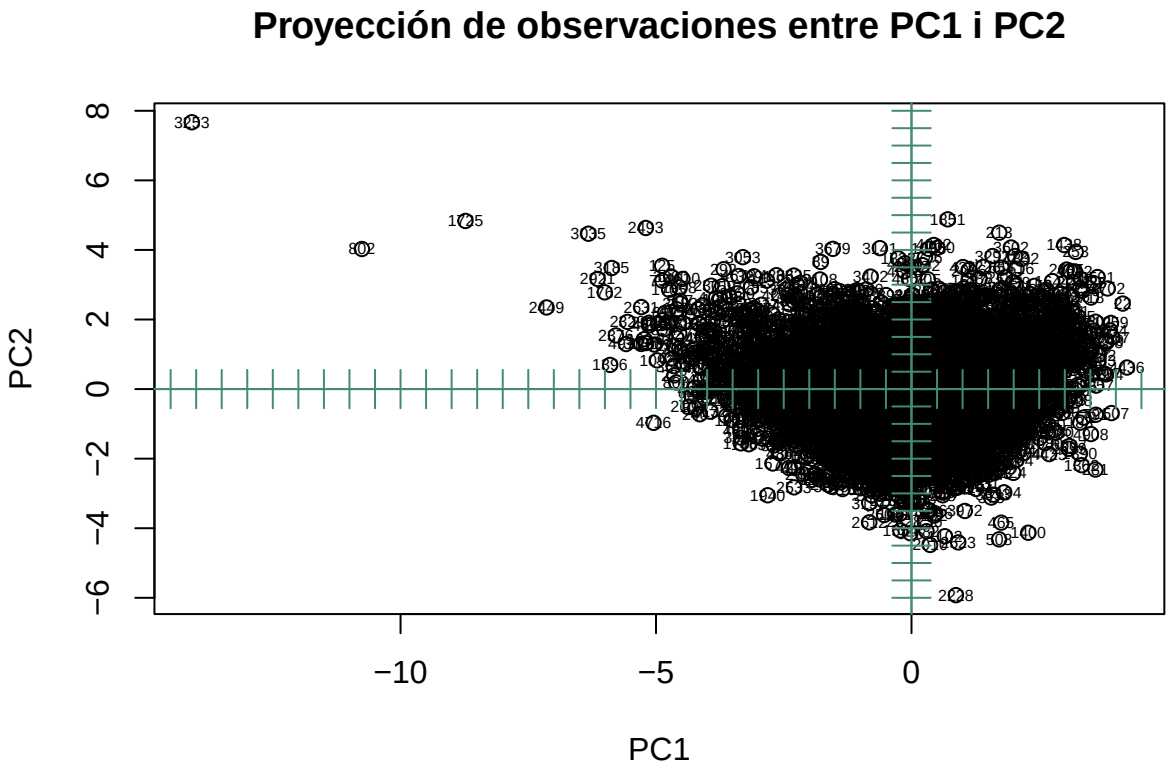
	PC1	PC2	PC3	PC4
OWN_CAR_AGE	0.0056870	0.3062623	-0.0030588	0.9203547
CNT_FAM_MEMBERS	-0.0247561	-0.0041456	-0.7074073	0.1177313
log_AMT_INCOME_TOTAL	0.1179538	-0.6836391	-0.0420849	0.1218891
log_AMT_CREDIT	-0.4904066	-0.4139790	-0.0598471	0.0681678
AGE_YEARS	-0.1154462	0.0462637	0.6823184	0.0564692
RATIO_CREDIT_INCOME	-0.5958824	0.1282325	-0.0355111	-0.0417412
RATIO_ANNUITY_CREDIT	0.3902375	0.3372441	-0.1098849	-0.2629810
DTI_RATIO	-0.4735547	0.3675972	-0.1237687	-0.2132899

- **PC1**: Las variables más fuertemente correlacionadas con esta dimensión son **RATIO_CREDIT_INCOME**, **log_AMT_CREDIT** y **DTI_RATIO**, todas correlacionadas de forma negativa y en este respectivo orden. Con ello, podemos pensar que el primer plano factorial (**PC1**) tiene relación con "**Nivel monetario según prestamos**". Puede entenderse que valores más elevados en la proyección sobre el primer plano factorial (**PC1**) indican individuos con unas diferencias menores entre el credito pedido y lo que ingresan anualmente, y con préstamos más bajos a nivel monetario.
- **PC2**: Las variables con mayor correlación con la segunda dimensión, en orden decreciente, son **log_AMT_INCOME_TOTAL** con correlación negativa, y **log_AMT_CREDIT** con correlación negativa y **DTI_RATIO** con correlación positiva. Se puede intuir que los individuos con valores más altos en la proyección del **PC2** serán aquellos con unos ingresos totales menores y creditos concedidos menores. Por lo tanto, el segundo plano factorial (**PC2**) podría quedar definido por "**Nivel de ingresos según créditos**"

- **PC3**: Para este tercer plano factorial, las variables más significativas son **CNT_FAM_MEMBERS** de forma negativa y **AGE_YEARS** de forma positiva. Así pues, aquellos individuos que cumplen estas características son clientes con familias poco numerosas y mayores (si su año de nacimiento es un valor alto, significa que son más mayores, dado a la correlación positiva con la variable de edad). Podría decirse que el tercer plano factorial (**PC3**) representa la "**Edad y grandaria familiar**".
- **PC4**: Para el cuarto plano factorial, se puede ver que la variable con mayor contribución en gran diferencia a las demás es **OWN_CAR_AGE**, correlacionada de forma negativa. Es decir, los clientes con valores de proyección en PC4 más grandes seran aquellos con coches más nuevos. Por lo tanto, el cuarto plano factorial (**PC4**) podría recibir el nombre de "**Edad vehículo** ".

Representación de individuos

A continuación se representan los individuos en los dos primeros planos factoriales.

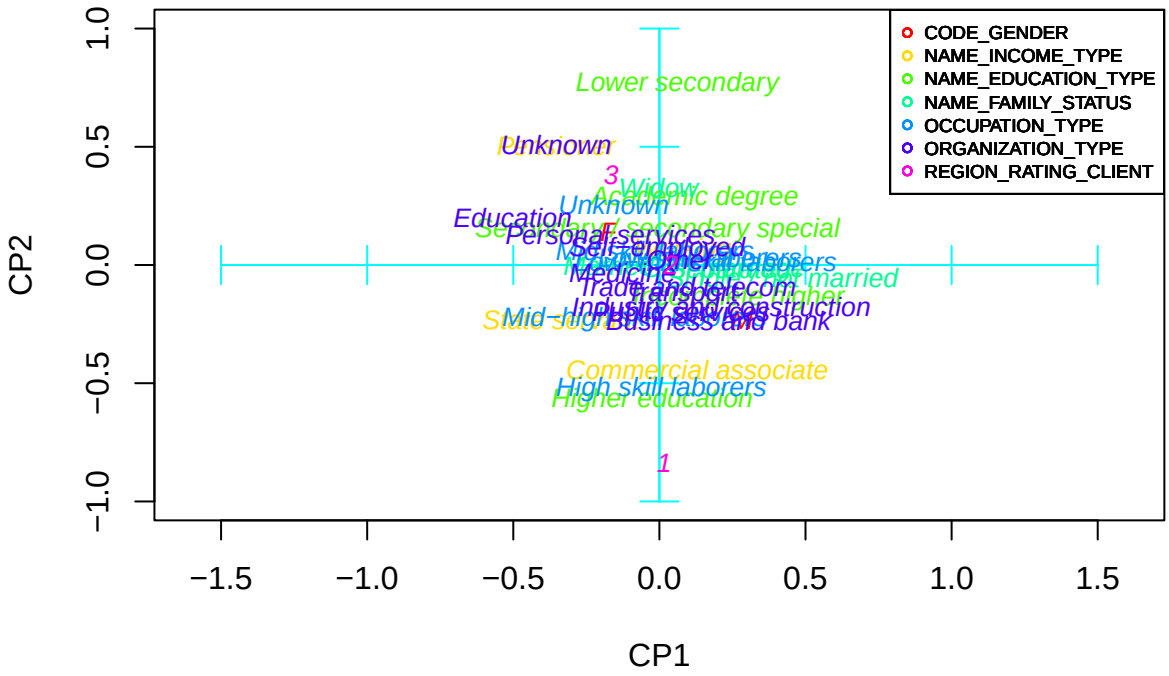


Como se puede observar, no se aprecian grupos diferenciados a partir de la proyección de los individuos. Hay una gran cantidad de estos que se concentran alrededor del origen de coordenadas, dando a entender que són individuos "ordinarios". Sí se observan algunos puntos alejados de la nube principal, estos perteneciendo a la representación de algunos individuos con características más extrañas a las del conjunto central de individuos.

Representación de variables categóricas en primeros planos factoriales

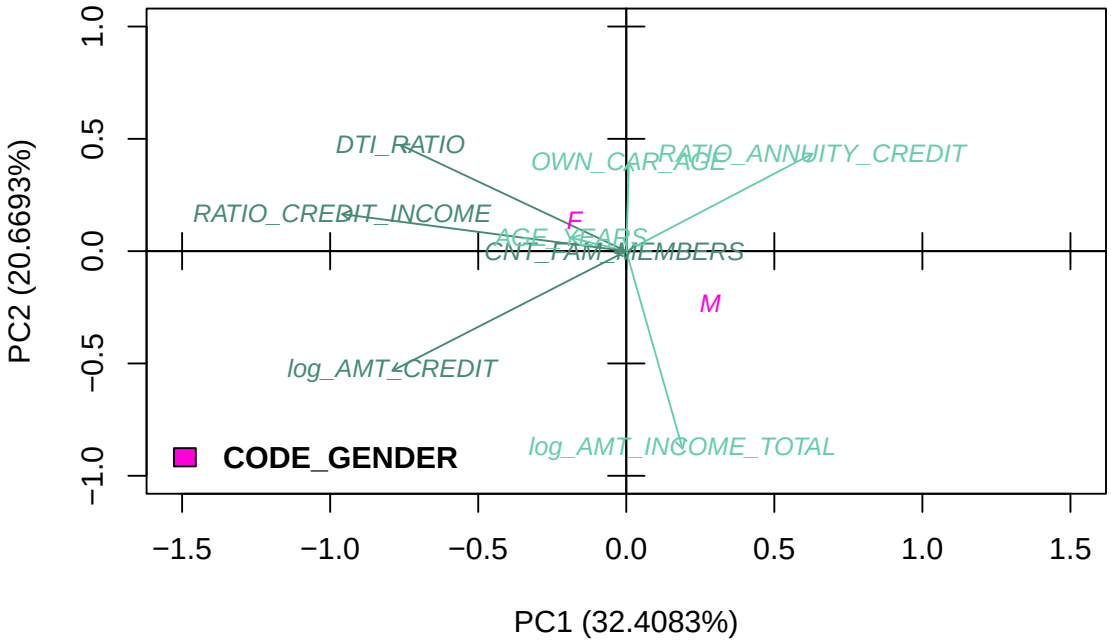
Una vez se han establecido los planos factoriales gracias a las variables nunéricas, es necesario representar también las variables categóricas para así acabar de hacer un estudio completo usando toda la base de datos

de la variable estudiada. De esta forma, se han representado los centroides de las coordenadas de cada nivel de cada variable categórica y se han obtenido los siguientes resultados:

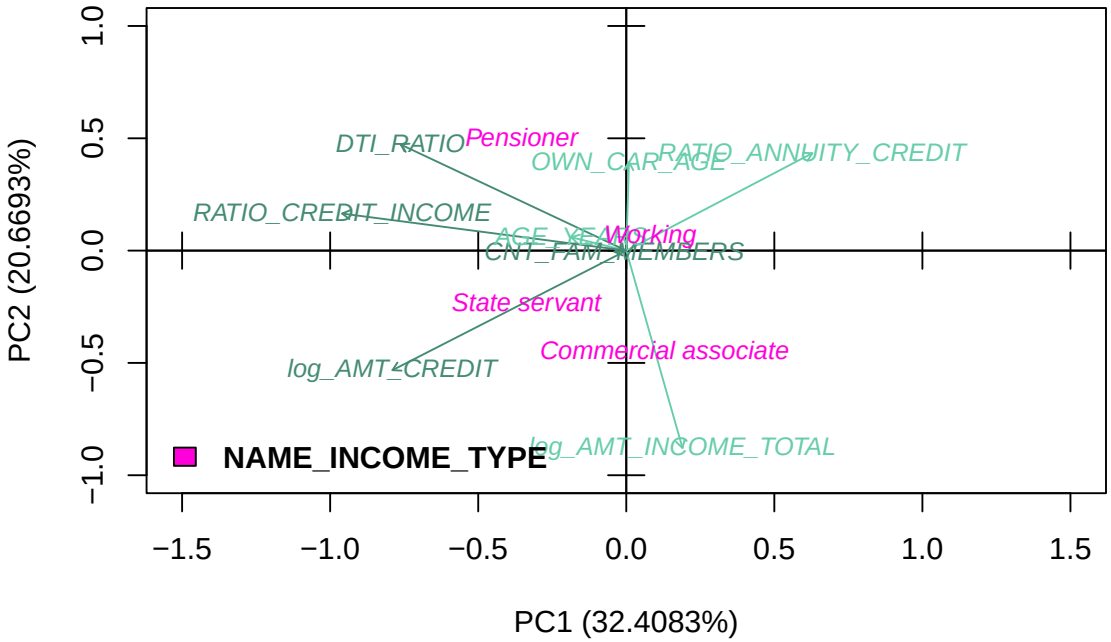


En este primer gráfico no se puede ver nada con claredad, por eso se ha decidido representar cada una de las variables categóricas en un gráfico distinto:

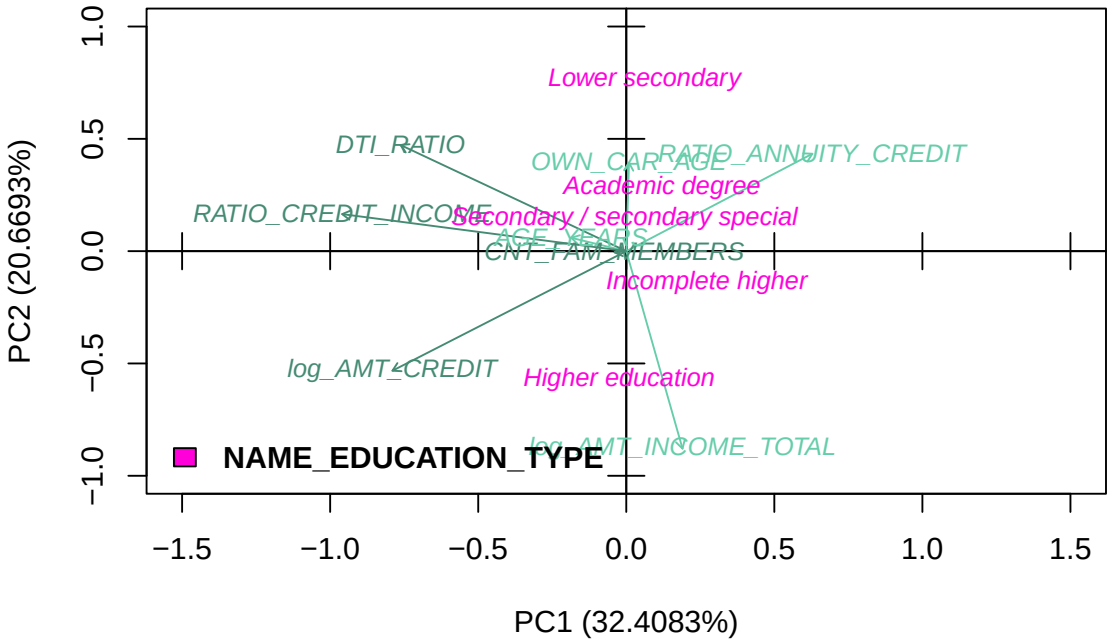
Proyecciones sobre el plano factorial de variables categóricas



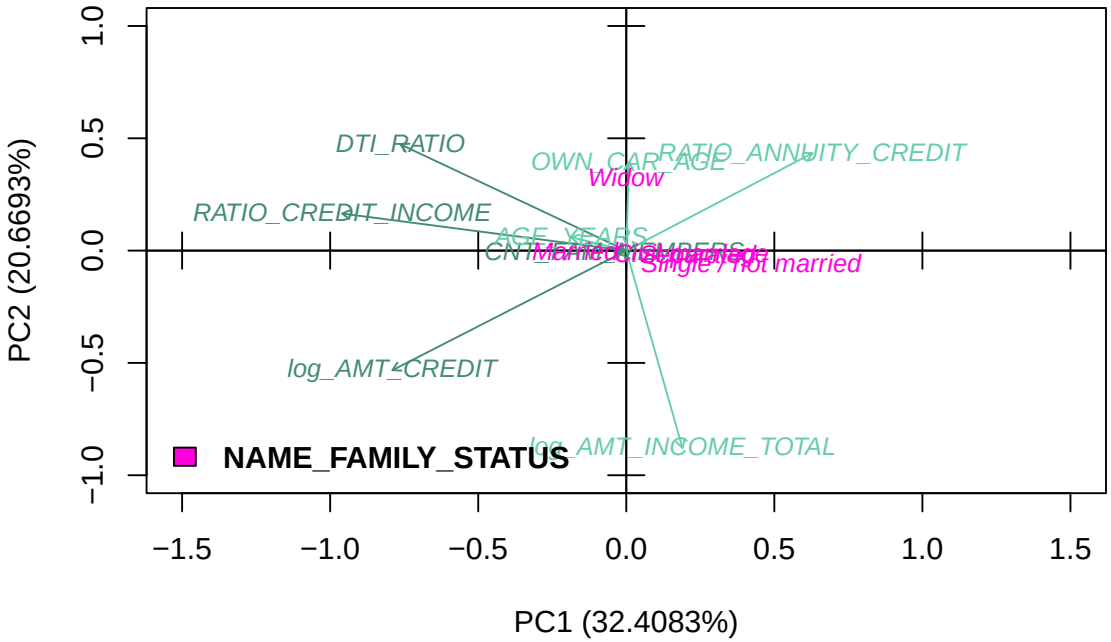
Proyecciones sobre el plano factorial de variables categóricas



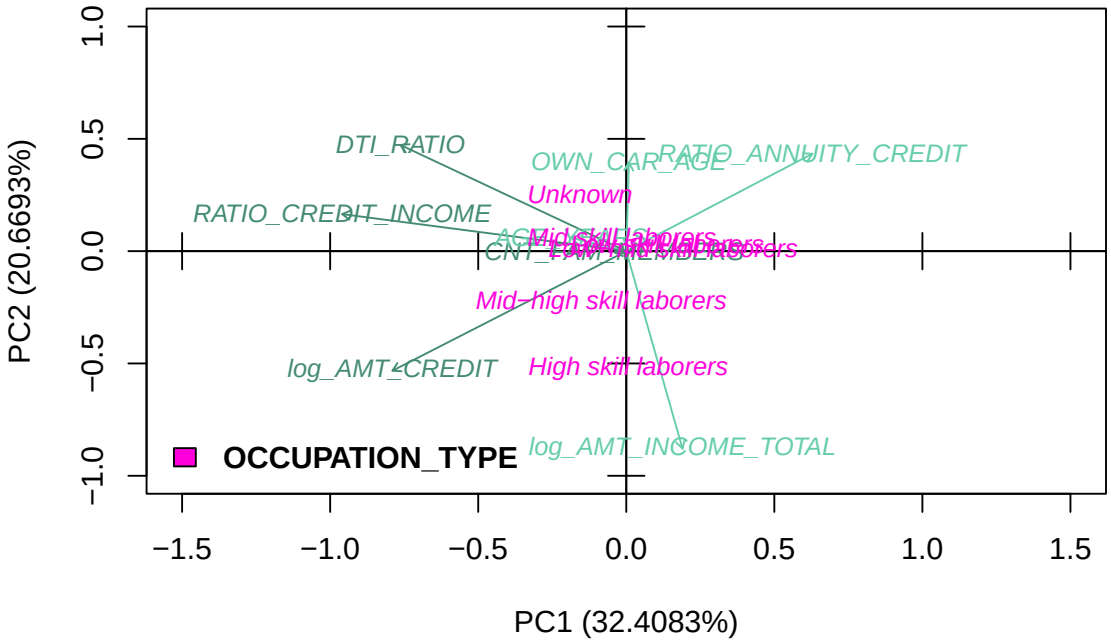
Proyecciones sobre el plano factorial de variables categóricas



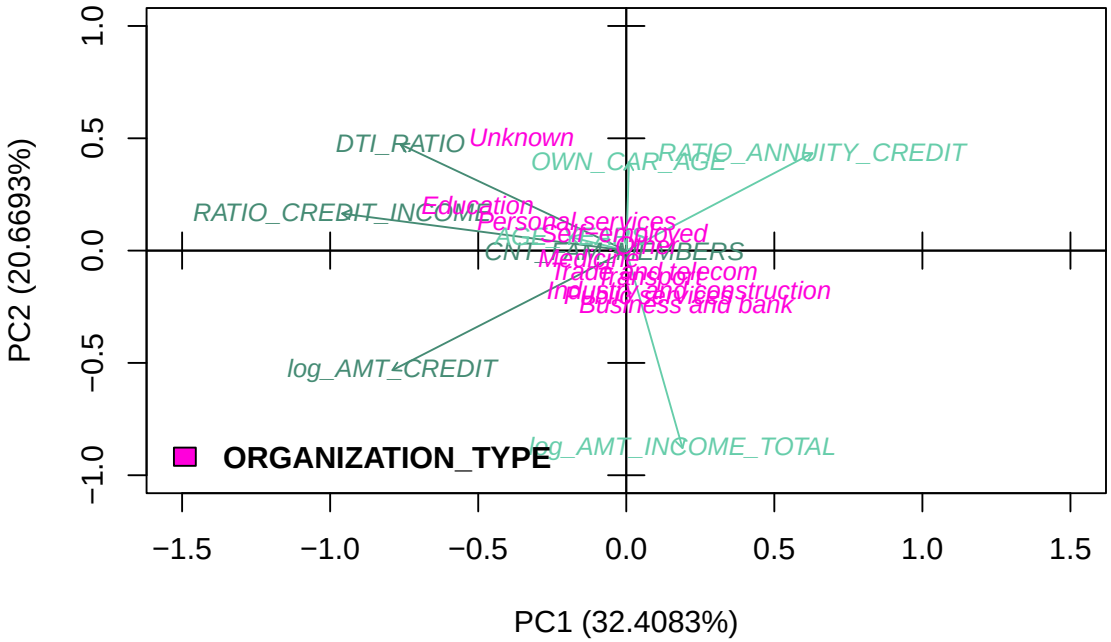
Proyecciones sobre el plano factorial de variables categóricas

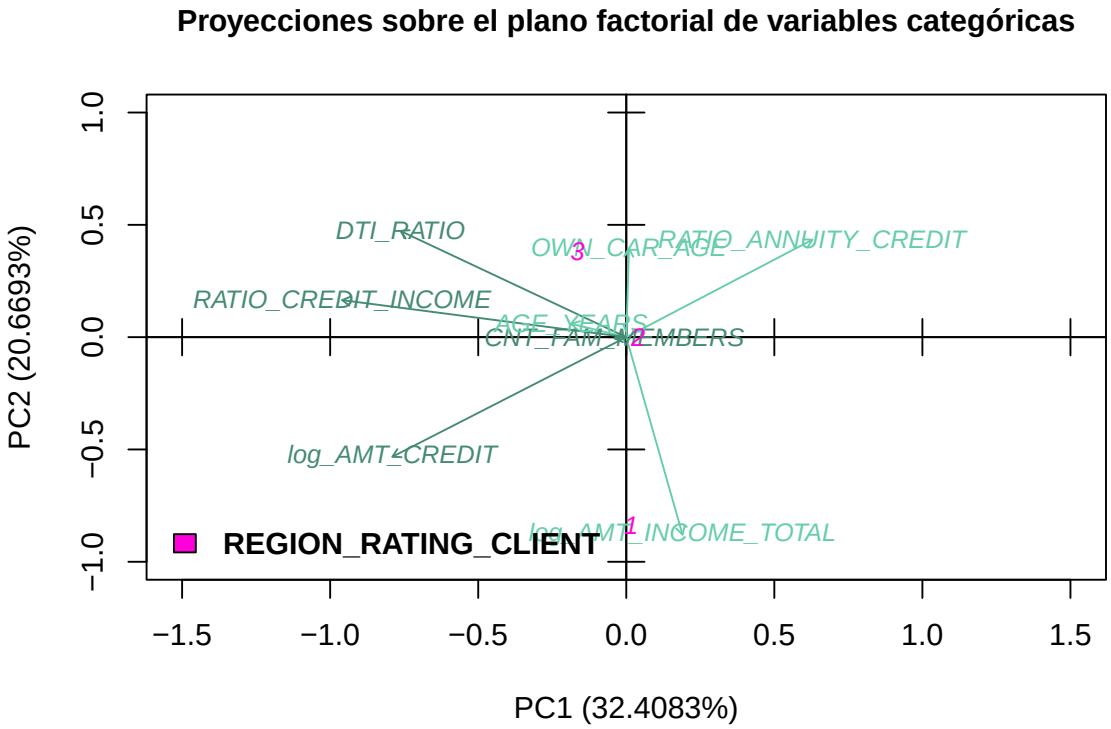


Proyecciones sobre el plano factorial de variables categóricas



Proyecciones sobre el plano factorial de variables categóricas





Algunos de los gráficos anteriores són interesantes de comentar. En el caso del gráfico que representa **NAME_EDUCATION_TYPE**, y de acuerdo con las descripciones establecidas de las dimensiones, se puede observar como los individuos con una educación "Lower secondary" són los que cuentan con unos ingresos totales menores y créditos concedidos de menor valor. Por otro lado, los individuos con una educación "Higher education" parecen ser los que piden crédito prestado de mayor valor monetario, y para los cuales sus ingresos totales son mayores. Uno de los motivos por los que se podria dar esto es por los préstamos solicitados para pagar la educación superior, y teniendo en cuenta que la base de datos es tomada en los Estados Unidos, se sabe que el precio de estos estudios es muy caro.

Observando el gráfico que incluye **CODE_GENDER**, se puede apreciar como los dos sexos presentan diferencias en la primera dimensión. De acuerdo con la explicación de la dimensión, los hombres són los que, en general, piden préstamos de menor valor monetario, y para los cuales la diferencia entre el crédito del préstamo y los ingresos anuales es menor. Es decir, que los hombres cuentan con menos años para pagar las deudas de los préstamos. Por el lado contrario, las mujeres presentan las características opuestas, préstamos más grandes y diferencias más significativas entre ingresos anuales y valor del crédito.

Por último, en el gráfico que representa la variable **NAME_INCOME_TYPE**, se pueden analizar las dos dimensiones por separado. Primero, si se comprueba la primera dimensión, es interesante ver que tanto los pensionistas como los funcionarios tienden a pedir préstamos de mayor valor, y ambos presentan diferencias entre ingresos anuales y el valor de dicho préstamo solicitado. Segundo, si se observa en función de la segunda dimensión, se aprecia que los pensionistas són los que presentan unos ingresos totales menores, mientras que las personas con ingresos derivados de puestos de trabajo comerciales són las que tienen ingresos totales mayores.