ACP

2023-11-01

```
##  CODE_GENDER             NAME_INCOME_TYPE
##  F:3098      Businessman         :   1
##  M:1902      Commercial associate:1111
##              Pensioner           : 763
##              State servant       : 306
##              Working             :2819
##
##
##                      NAME_EDUCATION_TYPE            NAME_FAMILY_STATUS
##  Academic degree              :   3     Civil marriage      : 546
##  Higher education             :1018     Married             :3095
##  Incomplete higher            : 156     Separated           : 320
##  Lower secondary              :  77     Single / not married: 798
##  Secondary / secondary special:3746     Widow               : 241
##
##
##                 OCCUPATION_TYPE                ORGANIZATION_TYPE
##  High skill laborers    : 660   Business and bank        :1505
##  Low-mid skill laborers :1047   Unknown                  : 763
##  Low skill laborers     : 722   Self-employed            : 708
##  Mid-high skill laborers: 432   Other                    : 390
##  Mid skill laborers     : 709   Industry and construction: 368
##  Unknown                :1430   Education                : 287
##                                 (Other)                  : 979
##  REGION_RATING_CLIENT TARGET   AMT_INCOME_TOTAL    AMT_CREDIT
##  1: 434               0:2865   Min.   :  29250   Min.   :  45000
##  2:3641               1:2135   1st Qu.: 112500   1st Qu.: 276278
##  3: 925                        Median : 135000   Median : 504000
##                                Mean   : 166849   Mean   : 578796
##                                3rd Qu.: 202500   3rd Qu.: 776402
##                                Max.   :2250000   Max.   :3375000
##
##   AMT_ANNUITY       DAYS_BIRTH      OWN_CAR_AGE    AMT_GOODS_PRICE
##  Min.   :  2673   Min.   :-25159   Min.   : 0.00   Min.   :  45000
##  1st Qu.: 16853   1st Qu.:-19130   1st Qu.: 5.00   1st Qu.: 234000
##  Median : 24876   Median :-15173   Median :10.00   Median : 450000
##  Mean   : 26831   Mean   :-15587   Mean   :11.12   Mean   : 515594
##  3rd Qu.: 33937   3rd Qu.:-11928   3rd Qu.:16.00   3rd Qu.: 675000
##  Max.   :177827   Max.   : -7711   Max.   :46.00   Max.   :3375000
##
##  CNT_FAM_MEMBERS log_AMT_INCOME_TOTAL log_AMT_CREDIT  log_AMT_ANNUITY
##  Min.   :1.000   Min.   :10.28        Min.   :10.71   Min.   : 7.891
##  1st Qu.:2.000   1st Qu.:11.63        1st Qu.:12.53   1st Qu.: 9.732
##  Median :2.000   Median :11.81        Median :13.13   Median :10.122
##  Mean   :2.166   Mean   :11.90        Mean   :13.05   Mean   :10.062
```

```
##  3rd Qu.:3.000   3rd Qu.:12.22        3rd Qu.:13.56   3rd Qu.:10.432
##  Max.   :8.000   Max.   :14.63        Max.   :15.03   Max.   :12.089
##
##  log_AMT_GOODS_PRICE   AGE_YEARS    DIFF_CREDIT_GOODS RATIO_CREDIT_INCOME
##  Min.   :10.71       Min.   :21.0   Min.   :-225000   Min.   : 0.125
##  1st Qu.:12.36       1st Qu.:32.0   1st Qu.:      0   1st Qu.: 2.032
##  Median :13.02       Median :41.0   Median :  47520   Median : 3.200
##  Mean   :12.93       Mean   :42.2   Mean   :  63201   Mean   : 3.901
##  3rd Qu.:13.42       3rd Qu.:52.0   3rd Qu.: 100980   3rd Qu.: 5.000
##  Max.   :15.03       Max.   :68.0   Max.   : 361746   Max.   :33.972
##
##  RATIO_ANNUITY_CREDIT   DTI_RATIO
##  Min.   :0.02528      Min.   :0.007514
##  1st Qu.:0.03799      1st Qu.:0.115149
##  Median :0.05000      Median :0.164963
##  Mean   :0.05418      Mean   :0.182008
##  3rd Qu.:0.06555      3rd Qu.:0.230162
##  Max.   :0.12003      Max.   :1.350750
##
```

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

PCA

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

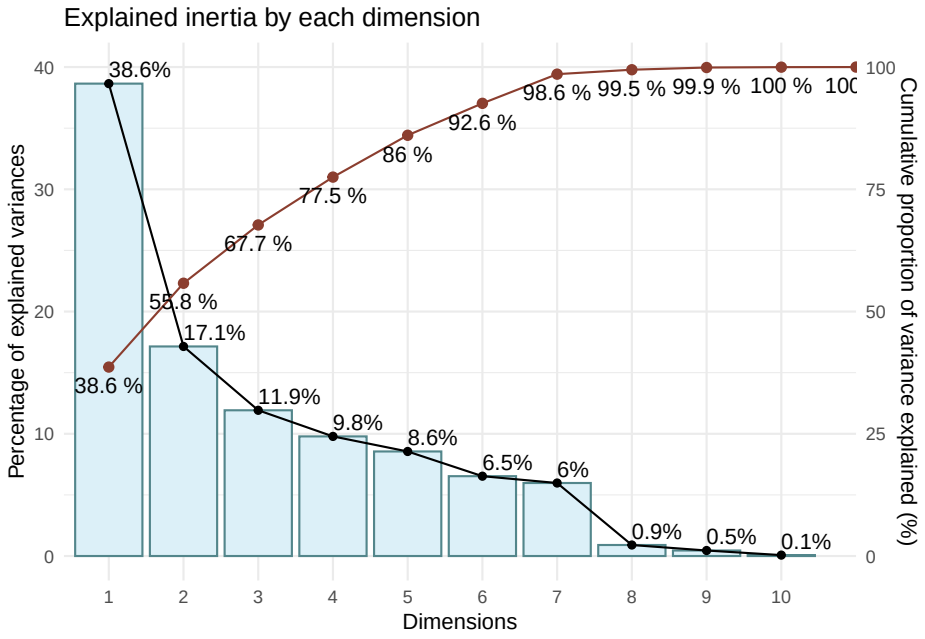

Figure 1: Inercia explicada por cada PF

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.

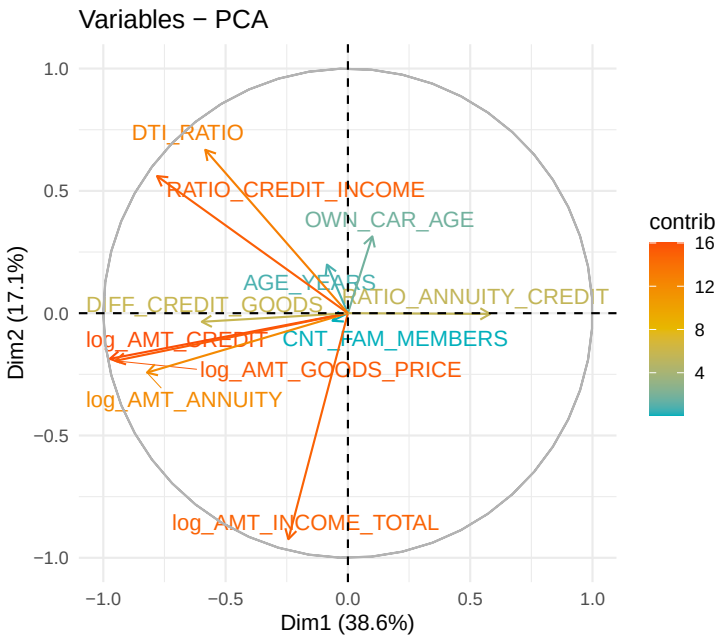

Figure 2: Gráfico de PF

Observamos la tabla de rotaciones:

Table 1: 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:

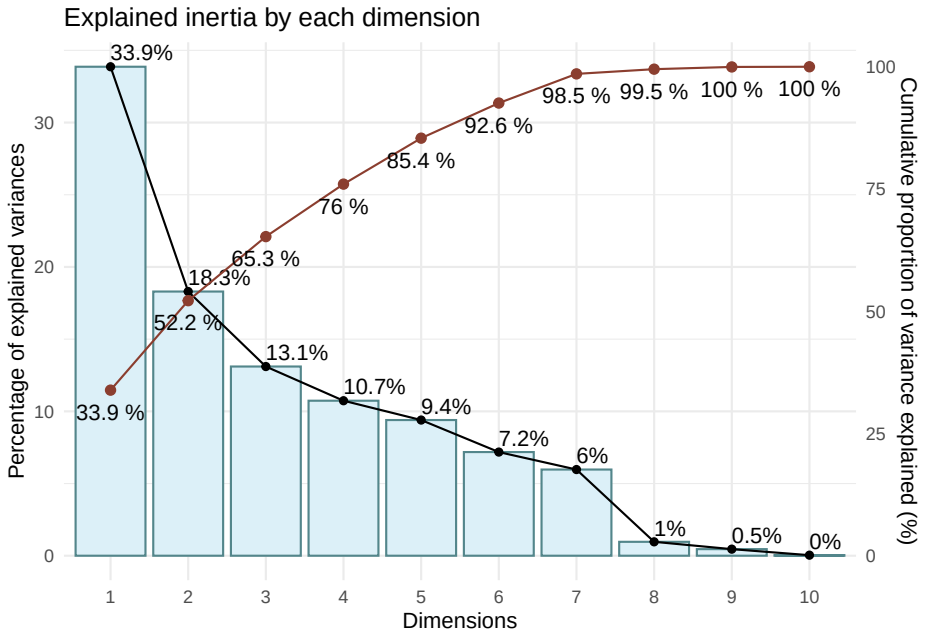

Figure 3: Inercia explicada por cada PF corregido

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 2: 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**.

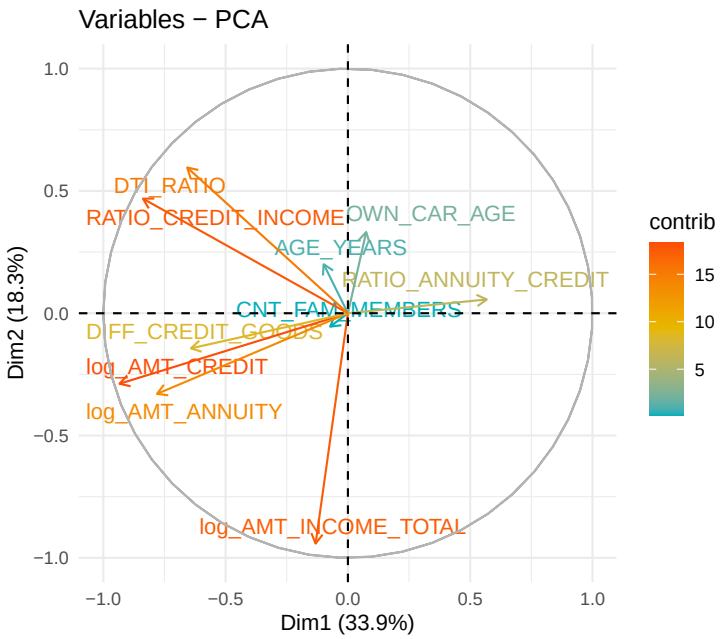

Figure 4: Gráfico de PF corregido

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 3: 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**.

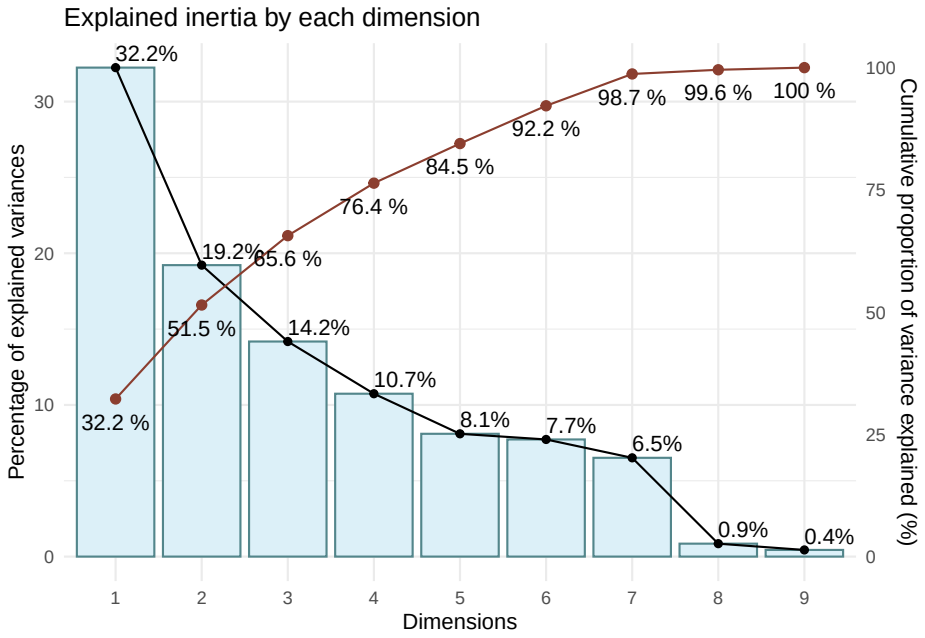

Figure 5: Inercia explicada por cada PF corregido

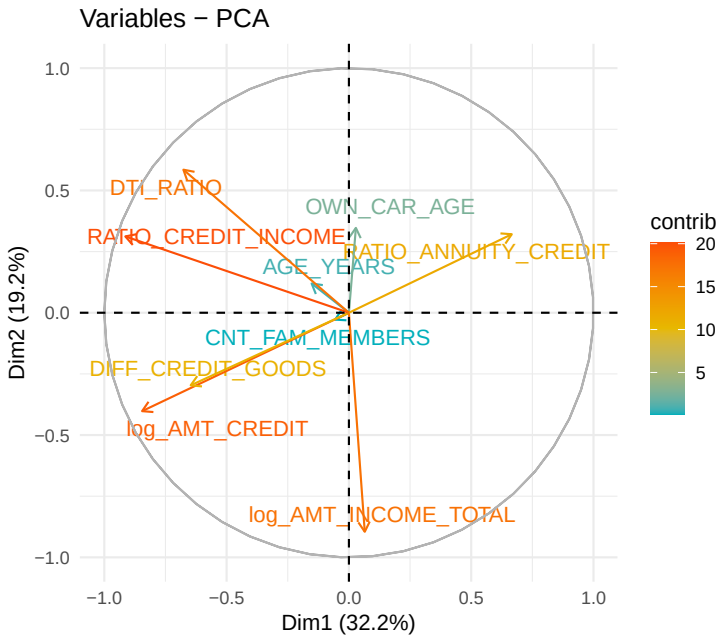

Figure 6: Gráfico de PF corregido

Ahora se conservan 8 variables numéricas.

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

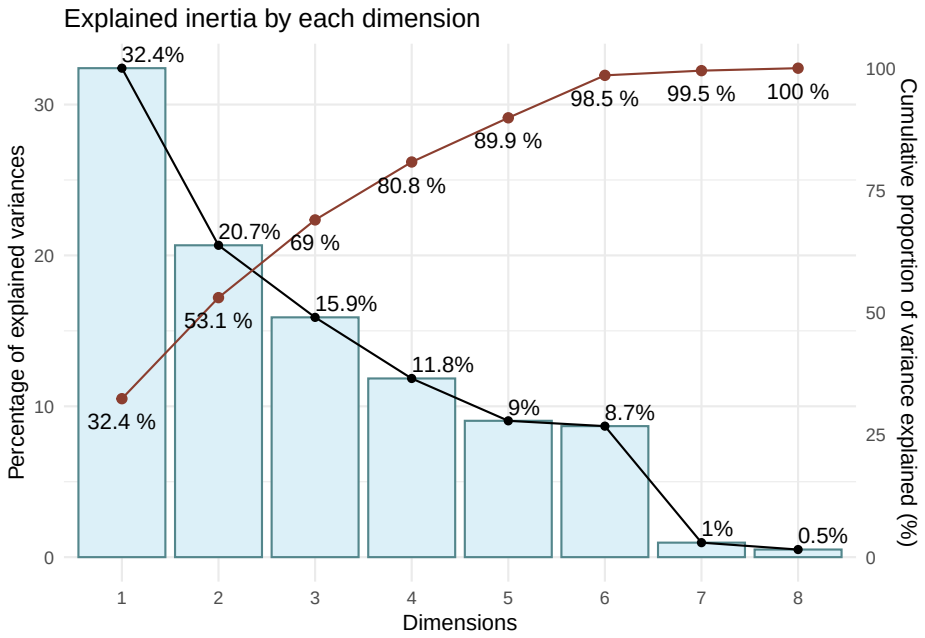

Figure 7: Inercia explicada por cada PF corregido

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 4: 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**

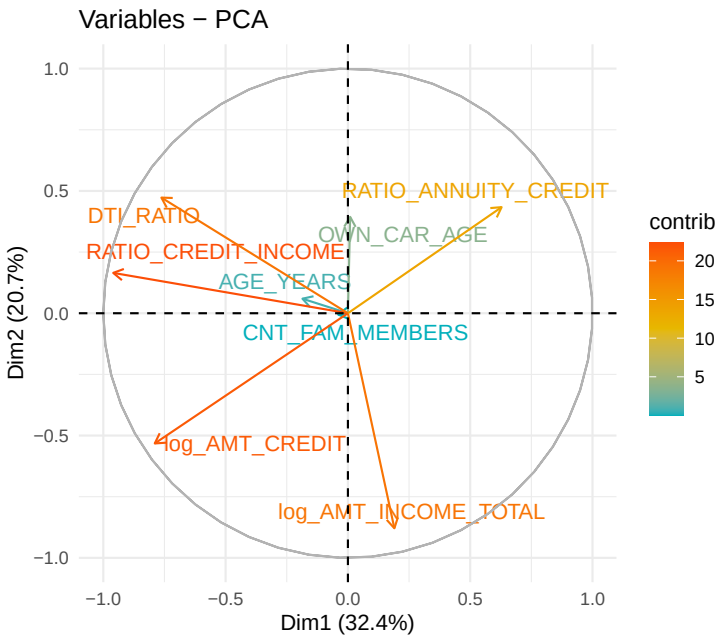

Figure 8: Gráfico de PF corregido

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.

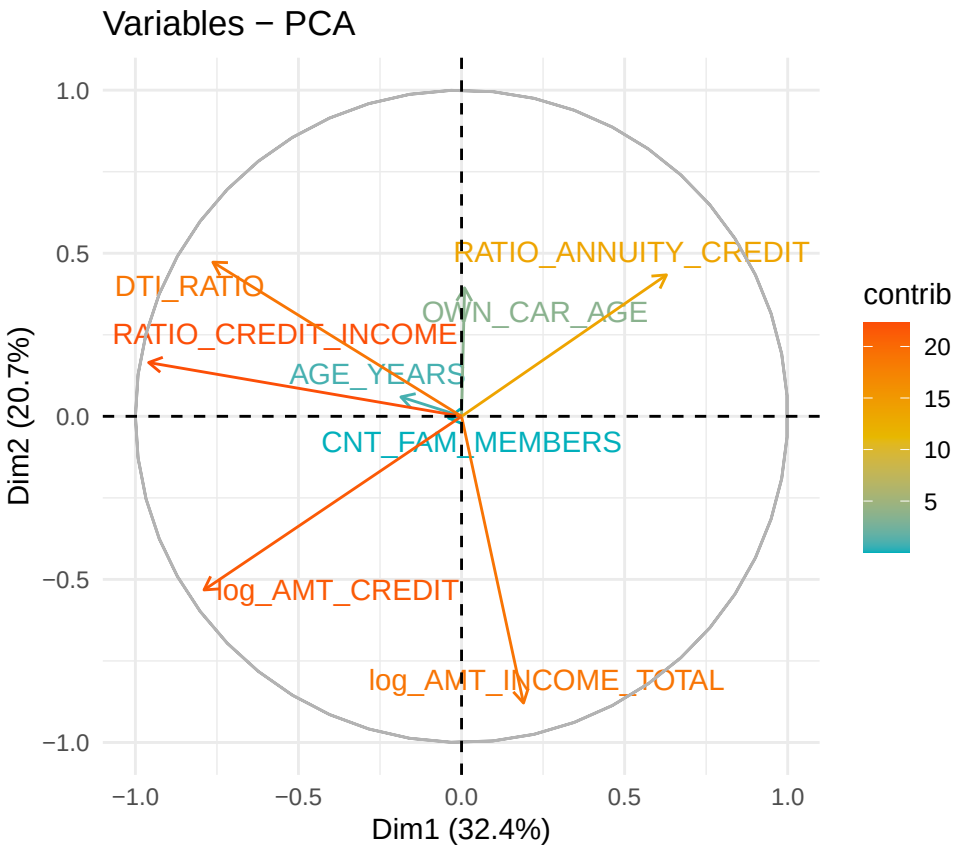

Table 5: Correlación de cada variable con cada plano factorial

	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

- **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 entre lo que pagan anualmente y lo que ingresan menor 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** ".