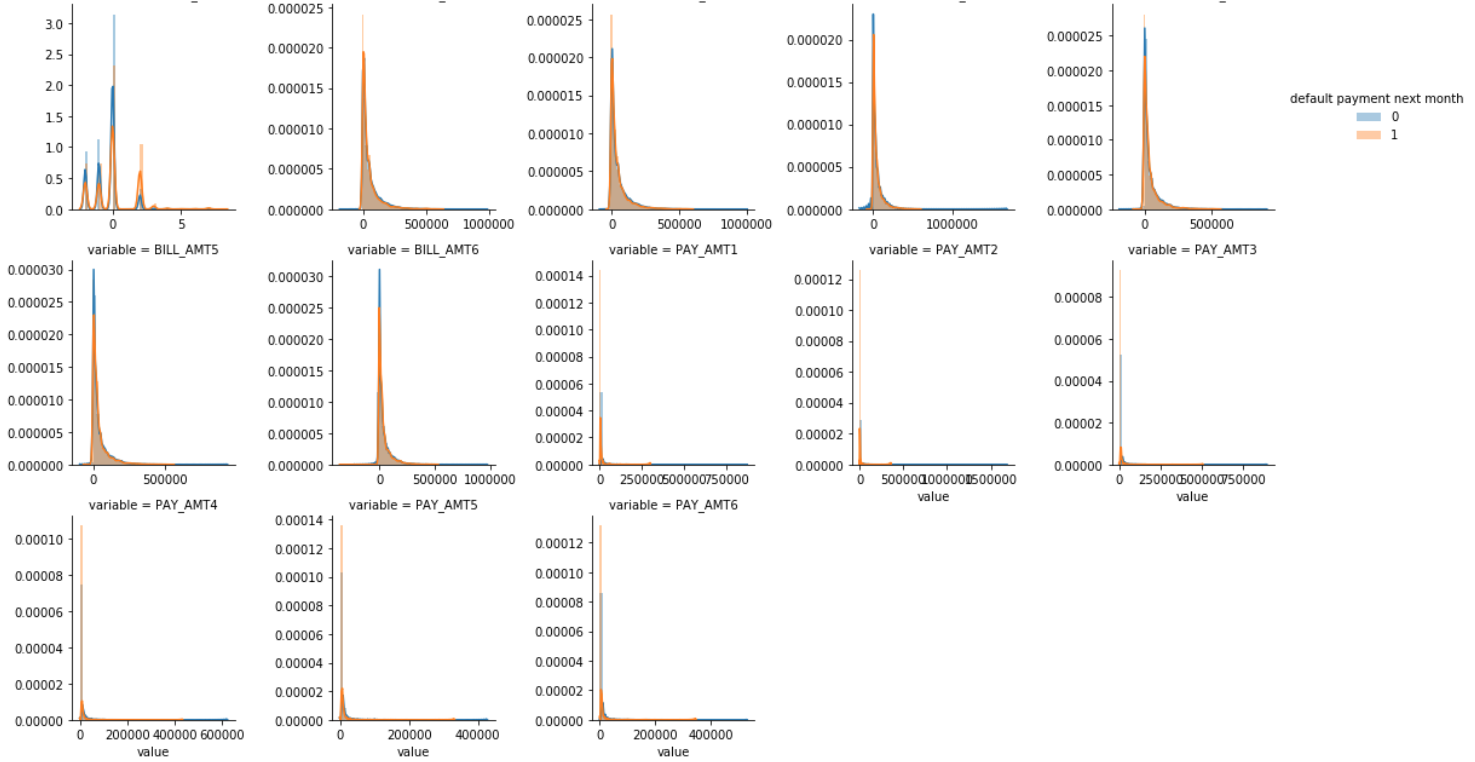
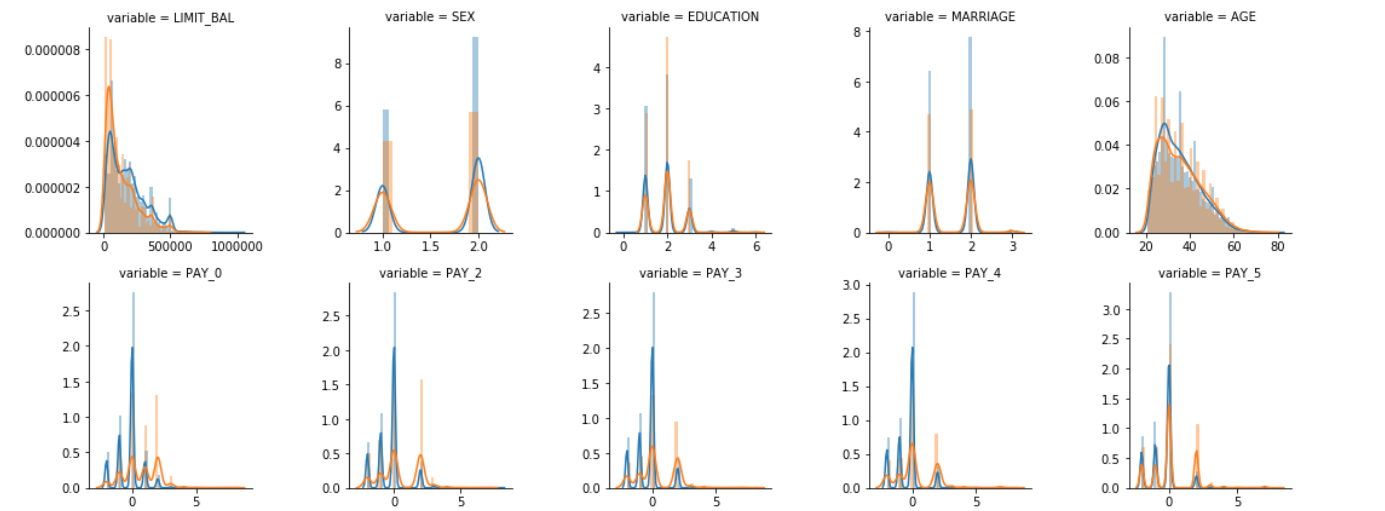
* Lecciones aprendidas:

Ya que nuestra variable dependiente es categórica, se puede dividir la muestra en resultados default/no default. Para esto realizamos diagramas para observar mejor cada distribución.



Observaciones:

* Default tienen una mayor proporción de valores LIMIT\_BAL inferiores.
* Los no default tienen una mayor proporción de mujeres (sexo = 2).
* Los no default tienen una mayor proporción de personas más educadas (EDUCATION = 1 o 2).
* Los no default tienen una mayor proporción de solteros (MARRIAGE = 2).
* Los no default tienen una mayor proporción de personas entre 30 y 40 años.
* Los no default tienen una proporción MUCHO mayor de variables PAY\_X cero o negativas, esto significa que estar al día o adelantarse a los pagos está asociado con no incumplir el mes siguiente. Esta es una relación fuerte ya que la distribución está más separada, por lo que esperamos que PAY\_X sea importante.

Como nota aclaratoria, se crea una función llamada ChiSquaredTestOfIndependence para determinar las relaciones entre dos variables nominales (categóricas). La frecuencia de cada categoría para una variable nominal es comparada a través de las categorías de la segunda variable nominal.

Con esta función evaluamos SEX, EDUCATION y MARRIAGE. Podemos observar lo siguiente:

* Default no es independiente de SEX ni MARRIAGE.
* EDUCATION no tiene todos los valores representados en default, esto nos lleva a analizar con mayor detenimiento los resultados para EDUCATION 0, 4, 5 y 6 que son los que no tienen valores suficientes para ser analizados. Se unen todos estos en una variable nueva: EDUCATION\_Corr y ejecutamos nuevamente la función para identificar que default tampoco es independiente EDUCATION\_Corr.

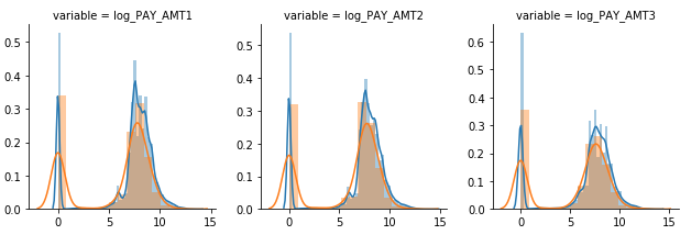
Se identifican las variables cuantitativas y cualitativas:

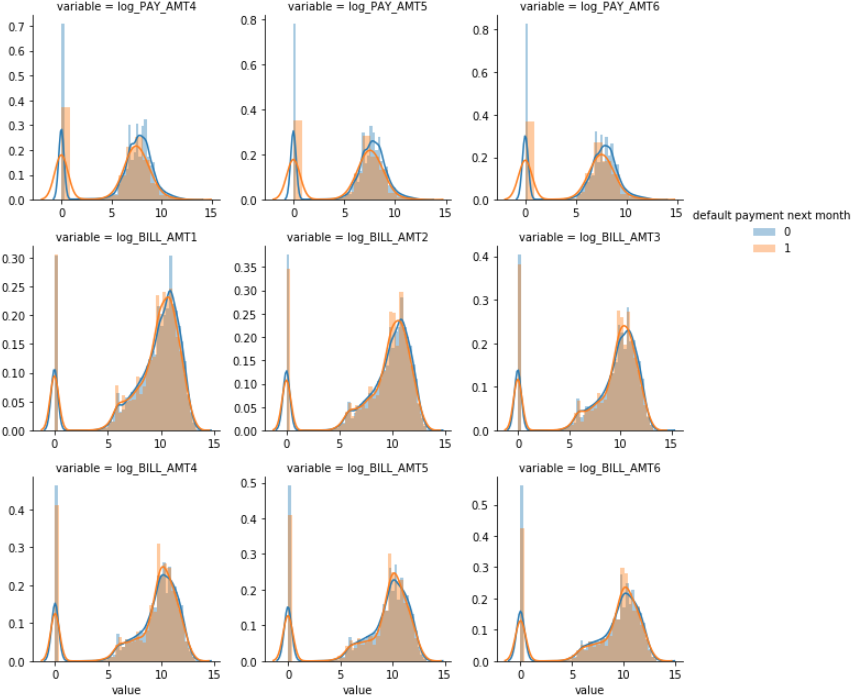
Cuantitativas:

* LIMIT\_BAL
* AGE

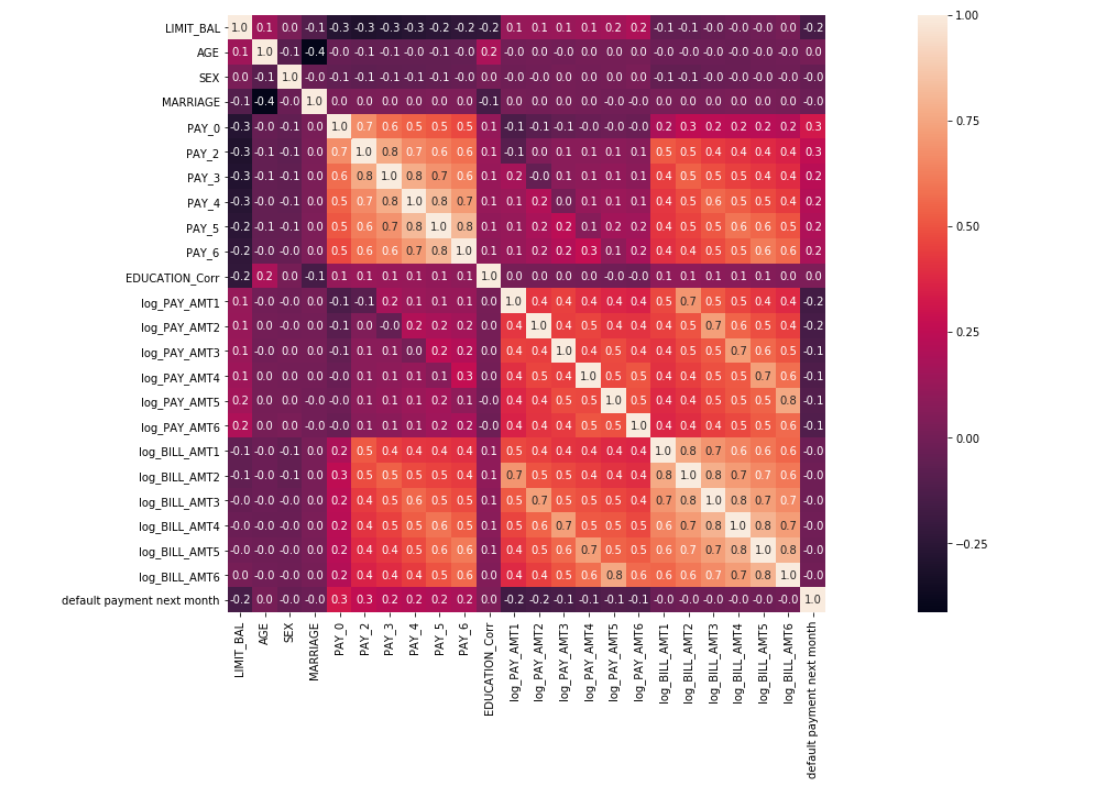
Cualitativas (todas las demás).

Con las variables PAY\_, podemos ver que son importantes, pero transformaremos las variables BILL\_AMT y PAY\_AMT de NT Dollars a Log (NT Dollars).





Verificamos las correlaciones con la variable default:



Por lo tanto, parece que las variables PAY\_0, PAY\_X son los predictores más fuertes de incumplimiento, seguidas por las variables LIMIT\_BAL y Log\_PAY\_AMT.