**Rapport demi parcours**

**Introduction**

L'objectif de ce projet est de réaliser une analyse quantitative du risque de défaut de paiement des cartes de crédit en en utilisant des modèles d'apprentissage automatique interprétables avec des données client accessibles, au lieu du score de crédit ou historique de crédit, dans le but d'aider et d'accélérer la prise de décision humaine

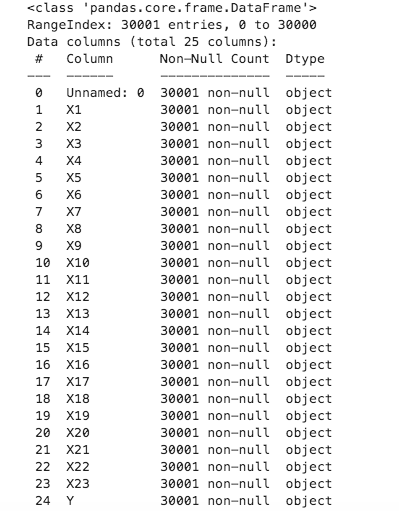
Pour l'analyse statistique dans ce projet de défaut de paiement par carte de crédit, j'ai suivi les étapes suivantes pour vérifier les résultats de l'analyse exploratoire graphique des données et identifier les corrélations entre les variables.

-Présentation de données

-Analyse des relations entre les données catégoriques

-Corrélation entre les différents variables.

**Présentation de données**



Notre base de données est composée de 30000 entrées dont chaque une est composée de 25 colonnes. Ces colonnes sont divisées comme suit:

-colonne ID (0)

-23 features (X1 -> X23) qui sont un mix de variables catégoriques et quantitatives.

Les variables catégoriques sont les variables SEX, EDUCATION, MARRIAGE et AGE.

Les variables quantitatives sont: LIMIT\_BAL, PAY\_0, PAY\_2, PAY\_3, PAY\_4, PAY\_5, PAY\_6,BILL\_AMT1, BILL\_AMT2, BILL\_AMT3, BILL\_AMT4, BILL\_AMT5, BILL\_AMT6, PAY\_AMT1, PAY\_AMT2, PAY\_AMT3, PAY\_AMT4, PAY\_AMT5, PAY\_AMT6,

-1 target (Y) qui est default payment next month.

**Explication des données:**

ID : ID de chaque client

LIMIT\_BAL : montant du crédit accordé en dollars NT (comprend les crédits individuels et familiaux/supplémentaires

SEXE : Genre (1=homme, 2=femme) ÉDUCATION : (1=école supérieure, 2=université, 3=lycée, 4=autres, 5=inconnu, 6=inconnu)

MARIAGE : État matrimonial (1=marié, 2=célibataire, 3=autres)

ÂGE : Âge en années

PAY\_1 : Statut de remboursement en septembre 2005 (-1=payer en bonne et due forme, 1=retard de paiement d'un mois, 2=retard de paiement de deux mois, ... 8=retard de paiement de huit mois, 9=retard de paiement de neuf mois et dessus) PAY\_2 : Statut de remboursement en août 2005 (échelle identique à celle ci-dessus) PAY\_3 : Statut de remboursement en juillet 2005 (échelle identique à celle ci-dessus)

PAY\_4 : Statut de remboursement en juin 2005 (échelle identique à celle ci-dessus) PAY\_5 : Statut de remboursement en mai 2005 (échelle identique à celle ci-dessus) PAY\_6 : Statut de remboursement en avril 2005

BILL\_AMT1 : Montant du relevé de facture en septembre 2005 (dollar NT) BILL\_AMT2 : Montant du relevé de facture en août 2005 (dollar NT)

BILL\_AMT3 : Montant du relevé de facture en juillet 2005 (dollar NT)

BILL\_AMT4 : Montant du relevé de facture en juin 2005 (dollar NT)

BILL\_AMT5 : Montant du relevé de facture en mai 2005 (dollar NT)

BILL\_AMT6 : Montant du relevé de facture en avril 2005 (dollar NT)

PAY\_AMT1 : montant du paiement précédent en septembre 2005 (dollar NT) PAY\_AMT2 : montant du paiement précédent en août 2005 (dollar NT)

PAY\_AMT3 : montant du paiement précédent en juillet 2005 (dollar NT)

PAY\_AMT4 : montant du paiement précédent en juin 2005 (dollar NT)

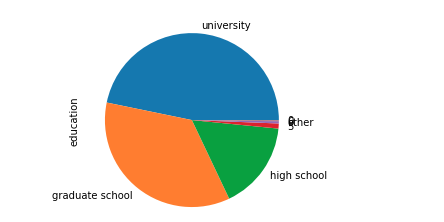
PAY\_AMT5 : montant du paiement précédent en mai 2005 (dollar NT)

PAY\_AMT6 : Montant du paiement précédent en avril 2005 (dollar NT)

default payment next month : Paiement par défaut (1=oui, 0=non)

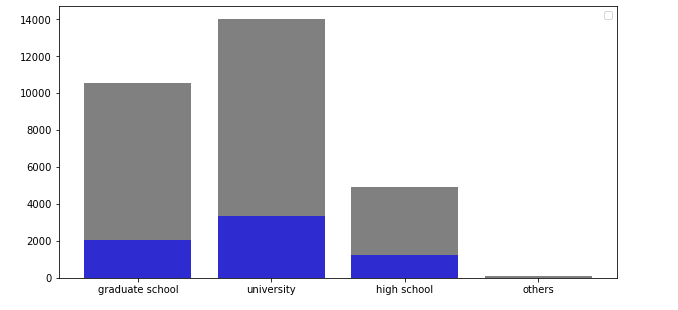
**Analyse de données**

**1- visualisation des valeurs d’éducation**

****

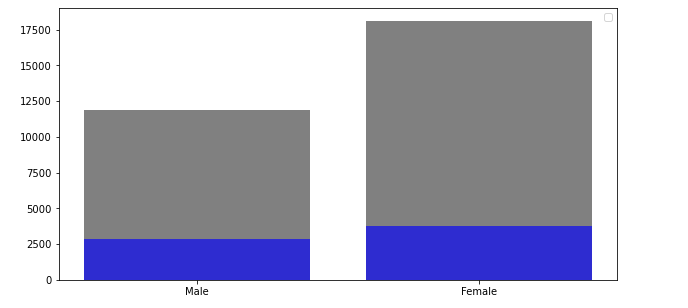
Ce graphe décrit la répartition de niveau d’étude en fonction de nombre individus dans la base de données. On remarque que le niveau universitaire est le niveau dominant de terme de nombre d’individus. On remarque aussi des niveaux dans la catégorie éducation qui n’ont aucune interprétation. Alors on a fini par les fusionner dans le niveau « Other »

**2- visualisation de paiement par défaut en fonction d’éducation**



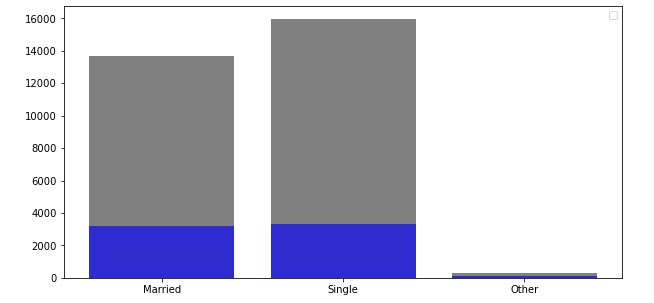
Cette figure décrit le nombre d’individus en fonction de leurs niveaux d’étude et leurs situations de paiement par défaut. Le graph en gris montre le nombre d’individus dans chaque niveau d’étude. Le graph en bleue montre le nombre d’individus dans chaque niveau d’étude qui ont une valeur de paiement par défaut de 1. D’après le calcul. 19% des « graduate school » ont une valeur de paiement par défaut de 1. 23% des « university » ont une valeur de paiement par défaut de 1. 25% des « high school» ont une valeur de paiement par défaut de 1. 5% des « university » ont une valeur de paiement par défaut de 1. On remarque que le pourcentage le plus élevé est chez la catégorie « high school » ce qui implique que l’éducation n’a pas un grand rôle pour la détermination de la valeur de paiement par défaut.

**3- visualisation de paiement par défaut en fonction du sexe**



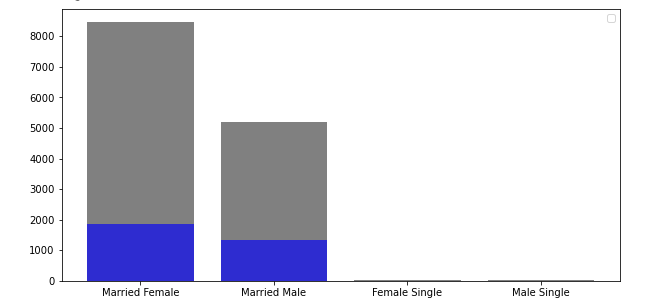
Cette figure décrit le nombre d’individus en fonction de leurs sexes et leurs situations de paiement par défaut. Le graph en gris montre le nombre d’individus de chaque sexe Le graph en bleue montre le nombre d’individus de chaque sexe qui ont une valeur de paiement par défaut de 1. D’après le calcul. 24.16% des hommes ont une valeur de paiement par défaut de 1. 20.77% des femmes ont une valeur de paiement par défaut de 1. On remarque que le pourcentage le plus élevé est chez la catégorie des hommes ce qui implique que les hommes ont plus de chance d’avoir la valeur de paiement par défaut de 1.

**4- visualisation de paiement par défaut en fonction du situation maritale**



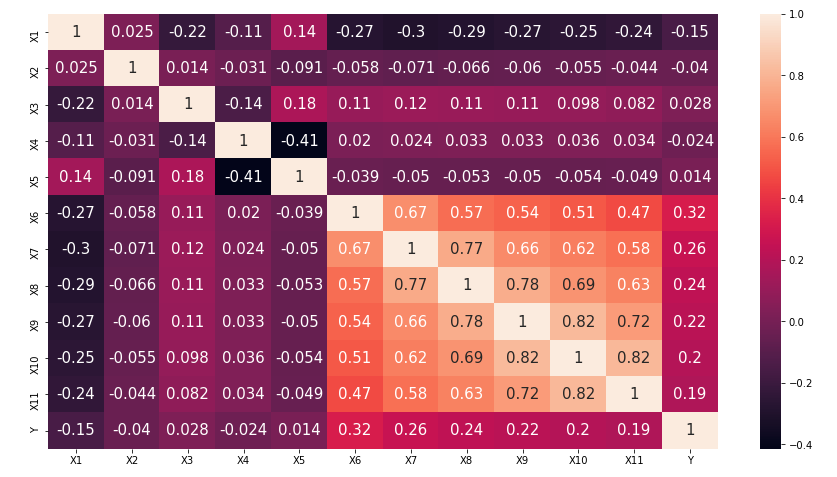
Cette figure décrit le nombre d’individus en fonction de leurs situations maritale et leurs situations de paiement par défaut. Le graph en gris montre le nombre d’individus dans chaque situation maritale. Le graph en bleue montre le nombre d’individus dans chaque situation maritale qui ont une valeur de paiement par défaut de 1. D’après le calcul. 23.47% des « married » ont une valeur de paiement par défaut de 1. 20.92% des « single» ont une valeur de paiement par défaut de 1. 26% des « Other» ont une valeur de paiement par défaut de 1. On remarque que le pourcentage le plus élevé est chez la catégorie « Other » Mais cela n’implique pas que cette catégorie a un grand rôle pour la détermination de la valeur de paiement par défaut.

**5- visualisation de paiement par défaut en fonction du sexe et de situation maritale**



Cette figure décrit le nombre d’individus en fonction de leurs situations maritale, leurs sexes et leurs situations de paiement par défaut. Le graph en gris montre le nombre d’individus de différent sexe et de situation maritale différente. Le graph en bleue montre le nombre d’individus de différent sexe et de situation maritale différente qui ont une valeur de paiement par défaut de 1. D’après le calcul. 21.96% des « Married Females» ont une valeur de paiement par défaut de 1. 25.93% des « Married Males» ont une valeur de paiement par défaut de 1. 7.5% des « Single Females » ont une valeur de paiement par défaut de 1. 14.28% des « Single Males » ont une valeur de paiement par défaut de 1. On remarque que le pourcentage le plus élevé est chez la catégorie « Married Males » Cela implique que les couples mariées ont plus de chance d’avoir de la valeur de paiement par défaut de 1.

**Correlation entre les variables**



Ce graphe est une application de la matrice de correlation. On Remarque que la variable X6 est fortement correlé avec les variables X7, X8, X9, et X10. On remarque aussi une forte correlation entre la variable X7 et X8, X9, et X10. Autrement dit. On peut éliminer les variables X7, X8, X9, et X10.