# Contexte de l'étude

# L'étude de cas concerne un directeur de banque qui s'inquiète du nombre croissant de résiliation dans le service des cartes de crédit. Dans une entreprise, le coût pour obtenir un nouveau client est généralement beaucoup plus élevé que ce qu'il faut pour conserver un client existant. À cette fin, l'objectif principal sera de prévoir le plus grand nombre de départs potentiels afin de permettre au directeur de proposer de manière proactive de meilleures offres aux clients.

**Info sur attrition/ rétention :**

[Qu'est-ce que le taux d'attrition et comment le réduire ?](https://www.saagie.com/fr/blog/definition-taux-attrition/)

# Fonctionnement des banques américaines

Réfléchir au ratio calcul d’indicateur calculé

Taux de surendettement?

# 

# Problématique

Prévenir les départs potentiels

# Import des librairies

# Import des données

# Description des données

L'ensemble de données se compose de 10127 observations, une pour chaque client (numéro de compte).

Il n’y a pas de doublon ni valeur manquantes.

Pour chaque compte, les informations pertinentes fournies sont les suivantes :

## La variable à expliquer

* **Attrition\_Flag**
  + Signification :Variable d'événement interne (activité du client)
  + Qualitative
  + Type : texte
  + 2 modalités :
    - Compte ouvert : Attributed Customer 16%
    - Compte fermé : Existing Customer 84%

## Les variables explicatives

### 5 variables qualitatives

#### Gender : Le genre Marital\_Status :Le statut marital Card\_Category : Le type de carte de crédits Income\_Category : le niveau de revenu Education\_Level : le niveau d'éducation

**14 variables quantitatives**

skim\_variable n\_missing complete\_rate mean sd p0 p25 p50 p75 p100

1 Customer\_Age 0 1 46.3 8.08 26 41 46 52 73

2 Dependent\_count 0 1 2.34 1.30 0 1 2 3 5

3 Months\_on\_book 0 1 35.9 8.02 13 31 36 40 56

4 Total\_Relationship\_Count 0 1 3.91 1.53 1 3 4 5 6

5 Months\_Inactive\_12\_mon 0 1 2.27 1.02 0 1 2 3 6

6 Contacts\_Count\_12\_mon 0 1 2.36 1.08 0 2 2 3 5

7 Credit\_Limit 0 1 8727. 9085. 1438. 2602 4644. 11253. 34516

8 Total\_Revolving\_Bal 0 1 1257. 758. 0 800 1364 1807 2517 9 Avg\_Open\_To\_Buy 0 1 7470. 9088. 15 1184. 3470. 9978. 34516

10 Total\_Amt\_Chng\_Q4\_Q1 0 1 0.773 0.218 0.256 0.643 0.743 0.86 3.40

11 Total\_Trans\_Amt 0 1 4655. 3513. 816 2385. 4100 4781. 18484

12 Total\_Trans\_Ct 0 1 68.7 22.9 11 54 71 82 139

13 Total\_Ct\_Chng\_Q4\_Q1 0 1 0.742 0.228 0.028 0.617 0.721 0.833 3.71

14 Avg\_Utilization\_Ratio 0 1 0.296 0.273 0 0.055 0.211 0.529 0.994

* **Dependent\_count** : Number of people who depend upon the customer for their support and welfare.
* **Months\_on\_book** : Period of relationship with bank in months
* **Total\_Relationship\_Count** : Total number of products held by the customer
* **Months\_Inactive\_12\_mon** : Number of months inactive in the last 12 months
* **Contacts\_Count\_12\_mon** : Number of contacts (phone calls) in the last 12 months
* **Credit\_Limit** : Credit limit on the credit card (Limite de crédit sur la carte de crédit)
* **Total\_Revolving\_Bal** : Total revolving balance on the credit card
* **Avg\_Open\_To\_Buy** : Open to buy credit line (average of last 12 months). This also turns out to be the difference between the credit limit (Credit\_Limit) assigned to a cardholder account and the present balance on the account (Total\_Revolving\_Bal).
* **Total\_Amt\_Chng\_Q4\_Q1** : Change in total transactions amount (Q4 over Q1)
* **Total\_Trans\_Amt** : Total transactions amount (last 12 months) Montant total des transactions (12 derniers mois)
* **Total\_Trans\_Ct** : Total number of transactions (last 12 months) Nombre total de transactions (12 derniers mois)
* **Total\_Ct\_Chng\_Q4\_Q1** : Change in total number of transactions (Q4 over Q1)*(Variation du nombre total de transactions (Q4 par rapport à Q1)*
* **Avg\_Utilization\_Ratio** : Average card utilization ratio (*Rapport d'utilisation moyen : Ratio moyen d'utilisation des cartes)*
* Education\_Level : Niveau d'éducation du client (lycée, diplôme, etc.)
* Marital\_Status : État civil du client (marié, célibataire, etc.)
* Income\_Category : Tranche de revenu du client en dollars (moins de 40K, 40K-60K, etc.)
* Card\_Category : Catégorie de carte de crédit (Blue, Silver, etc.)
* Months\_on\_book : Période de relation avec la banque en mois
* Compte\_Relations\_Total : Nombre total de produits détenus par le client
* Months\_Inactive\_12\_mon : Nombre de mois d'inactivité au cours des 12 derniers mois
* Contacts\_Count\_12\_mon : Nombre de contacts (appels téléphoniques) au cours des 12 derniers mois
* Credit\_Limit : Limite de crédit sur la carte de crédit
* Total\_Revolving\_Bal : Solde total renouvelable de la carte de crédit
* Avg\_Open\_To\_Buy : Ligne de crédit ouverte pour l'achat (moyenne des 12 derniers mois). Il s'agit également de la différence entre la limite de crédit (Credit\_Limit) attribuée au compte du titulaire de la carte et le solde actuel du compte (Total\_Revolving\_Bal).
* Total\_Amt\_Chng\_Q4\_Q1 : Variation du montant total des transactions (Q4 sur Q1)
* Total\_Trans\_Amt : Montant total des transactions (12 derniers mois)
* Total\_Trans\_Ct : Nombre total de transactions (12 derniers mois)
* Total\_Ct\_Chng\_Q4\_Q1 : Variation du nombre total de transactions (Q4 par rapport à Q1)

# Statistiques descriptives

Les stat desc sont réalisées sur les deux populations de notre base de données :

Data\_stay :clients qui restent cad qui ont un flag\_attrition =0 8500 observation

et Data\_quit : clients ayant résilié leur contrat, leur flag\_attrition =1 1627 observations

On a 10 127 observations au total sur le dataset de départ.

Comparaison de moyenne  
(CF) Test de student, test de wilcoxon….

RB Difficultés avec lecture slides 10 (graphe shapirowilk) et résultats tests student et wilcoxon

# Création d’indicateur calculés

(CF) Réfléchir au ratio : calcul d’indicateur calculé

Taux de surendettement?

# Matrice de corrélation

OK

# AFC/ACM

(LH) Analyse factorielle des **correspondances** entre 2 variables quantitatives

# CLUSTERS (K Means)

(RB) Définition de groupes de population et établir un modèle par groupe de population

# 

# Sélection des variables => recherche du meilleur modèle

## Mettre en facteur toute les variables

## 

## Créer un échantillon de même effectifs pour chacunes des modalités de la variables à expliquer pour les test

* **Séparer cet échantillon**
  + - Une partie pour le training (⅔)
    - Une partie pour les test(⅓) ( = “ noter le modèle")

=> /!\ évite le sur-aprentissage

## Par iteration

Pour deux modèles au résultats identiques => aller vers le plus simple (le moins de variables)

## 

**Modele par regression logistique**

.