Traitement des valeurs aberrantes:

- Elimination des instances
- Remplacement par valeurs IQRMin , IQRMax , Médiane, Mode
- Discrétisation.
- Laisser tel qu'ils sont car données réelles pertinentes.
- Remplacer la valeur aberrante par NaN ou Null.
- Régression
    - **MonthlyIncome** **[3]** **k= 1+ (10/3)*$\log_{10}$(N)**
    - **NumCompaniesWorked** **[4]**
    - **PerformanceRating** **[2]** **< Médiane >**
    - **StockOptionLevel** **[2]** < **Médiane** >
    - **TotalWorkingYears** **[3]** **k= 1+ (10/3)*$\log_{10}$(N)**
    - **TrainingTimesLastYear** **[4]**
    - **YearsAtCompany** **[4]**
    - **YearsInCurrentRole** **[4]**
    - **YearsSinceLastPromotion** **[4]**
    - **YearsWithCurrentManager** **[2]** **< IQRMin , IQRMax >**

Traitement des valeurs manquantes:

N	Option de traitement	Type	Sym ?	Autre
1	supprimer la ligne	-	-	~ perte d'information
2	supprimer la colonne	-	-	Plus de 50% de valeurs manquantes
3	remplacer par la moyenne	Numérique	oui	
4	remplacer par la médiane	Numérique	all	
5	remplacer par le Mode	all	all	~ biaiser les données parfois
6	remplacer par la moyenne par classe	Numérique	oui	
7	remplacer par la médiane par classe	Numérique	all	
8	remplacer par le Mode par classe	all	all	
9	remplacer par la valeur suivante ou précédente.	-	-	time series data
10	recherche google / expert	all	all	si disponible
11	nouvelle valeur signifiant "unknown" ou une Constante	all	all	dépend de la sémantique
12	méthodes avancées (clustering, arbre de décision,..)	all	all	

- **EnvironmentSatisfaction** <Asym, numérique, catégorique> **[8]**
- **MonthlyIncome** <Asym, numérique> **[7]**

Codification des données textuelle catégorique ordinales et binaires

- **BusinessTravel [0, 1, 2]**
- **Attiration [0, 1]**
- **Gender [0, 1]**

Réduction de la dimensionnalité:

- Elimination des attributs à faible variance (même valeur partout)
  - **EmployeeCount**
  - **Over18**
  - **StandardHours**
  - **PerformanceRating**
- Elimination des attributs non pertinents
  - **EmployeeNumber**
- Elimination des attributs redondants (corrélation/relation)
  - **RAS**
- Elimination des instances répétées.
  - **RAS**

Discrétisation Lissage des données (aberrantes) :

- **par intervalles égaux**
- ~~par fréquences égales~~

Normalisation/stanrdadization:

- **par min-max**
- ~~par z-score~~

Ordre des différentes étapes de pré-traitement :

Discrétisation peut causer des redondances entre les instances
⇒ Discrétisation **PUIS** Réduction de la dimensionnalité H (éliminer une instance)
Traitement des outliers peut conduire à une faible variance d'un attribut
⇒ Traitement des outliers **PUIS** Réduction de la dimensionnalité V (éliminer un attribut) ?
L'existence d'outliers peut fausser la normalisation
⇒ Traitement des outliers **PUIS** Normalisation ?

Etapes:

1- Codification 2- Traitement des Outliers 3- Traitement des valeurs nulles 4- Discrétisation **5- Réduction Dimensionnalité V** **6- Réduction Dimensionnalité H** 7- Normalisation	1- Codification 2- Traitement des Outliers 3- Traitement des valeurs nulles **4- Réduction Dimensionnalité V** 5- Discrétisation **6- Réduction Dimensionnalité H** 7- Normalisation