République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université des Sciences et de la Technologie Houari Boumédiène

Faculté d'Informatique
Département d'Intelligence Artificielle

Master 2 Systèmes Informatiques intelligents

Module : Data Mining

Rapport de projet, partie 2
Prétraitement des données

Réalisé par :
BOUROUINA Rania, 181831052716
CHIBANE Ilies, 181831072041

Année universitaire : 2022 / 2023

Table des matières

Introduction Générale

Le Data Mining est une technique qui consiste à examiner une grande structure de données pour trouver des modèles, des tendances, des idées cachées qui ne seraient pas possibles en utilisant des techniques plus simples basées sur des requêtes. Cette technique utilise des algorithmes mathématiques sophistiqués pour classer, diviser, segmenter l'ensemble des données, les pré-traiter si nécessaire et évaluer la possibilité d'événements futurs.

Dans ce projet, nous allons mettre en pratique toutes les notions vues dans le cours en passant par plusieurs phases :
1. Analyse des données
2. **Prétraitement des données**
3. Extraction de motifs fréquents, règles d'associations et corrélations
4. Classification et Prédiction
5. Clustering

Lors de la première partie, nous avions constaté de nombreux problèmes récurrents au sein du dataset. Que ce soit les données manquantes ou aberrantes, ou certains attributs inutiles ou problématiques. C'est pour cela que lors de cette deuxième partie, nous allons mettre en application différentes méthodes afin de remédier à ces différentes contraintes ainsi que de réaliser différent autre prétraitement afin d'améliorer et d'optimiser notre dataset. Nous passerons donc en revu les méthodes de traitements des données manquantes et aberrantes, la réduction de données verticale et horizontale ainsi que via la discrétisation, et pour finir, nous présenterons différentes méthodes de normalisation. Nous conclurons ce rapport en présentant l'interface réalisée afin de réaliser le prétraitement via le GUI.

Traitement des valeurs manquantes et aberrantes

Introduction Les valeurs manquantes et aberrantes sont des contraintes courantes rencontrées lors de la phase d'analyse des données. Grandement handicapante lors des différentes opérations effectuées nécessitant le dataset, il est impératif d'y remédié au plus tôt. C'est pour cela que dans ce chapitre, nous nous attardons sur différentes méthodes permettant de les gérer.

1.1 Traitement des valeurs manquantes

Les valeurs manquantes, comme son nom l'indique, représente les valeurs NaN d'un attribut. ces valeurs se doivent d'être comblées afin d'éviter tous problèmes futurs. Dans le cas de notre dataset, nous avons les valeurs manquantes suivantes :

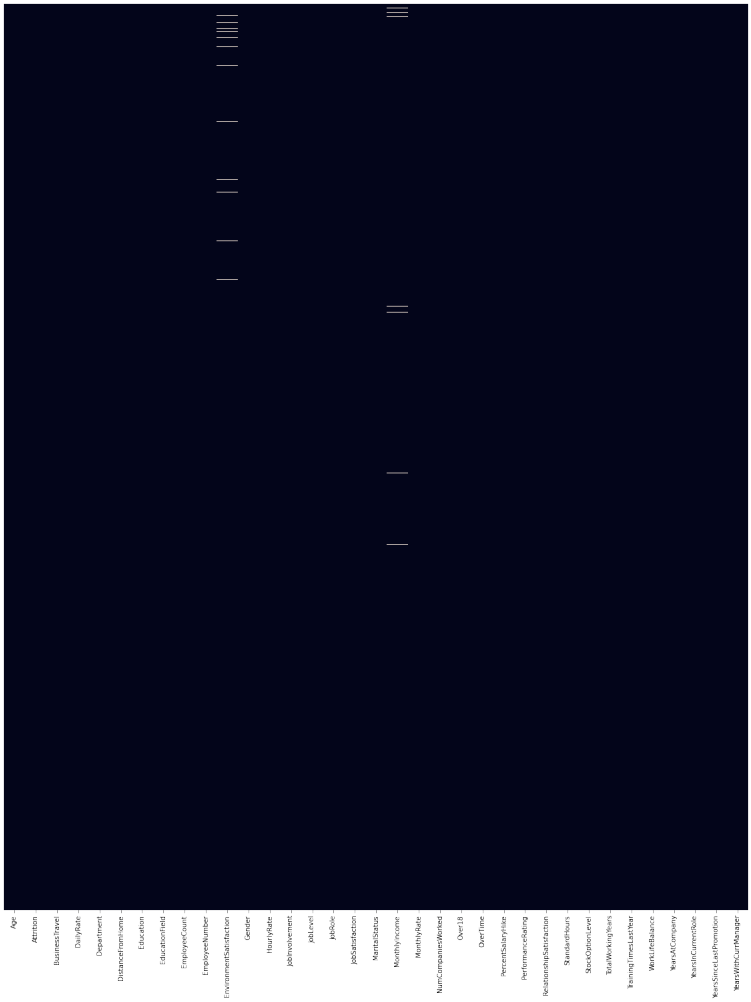


FIGURE 1.1 – Aperçu du Dataset

Comme on peut le constater, deux attributs sont concernés, EnvironmentSatisfaction et MonthlyIncome. Les moyens de remplacer ces valeurs sont nombreuses et nous allons les détailler une par une.

1.1.1 En utilisant les mesures de tendances

L'une des méthodes la plus utilisée et la plus simple et de remplacer les valeurs par l'une des mesures de tendances. À commencer par la moyenne qui comme vu dans la partie 1 se calcule ainsi :

$$\bar{X} = \frac{(X_1 + X_2 + \cdots \ldots + X_n)}{(n)} = \frac{1 \sum X_j}{n}$$

FIGURE 1.2 – Description globale du dataset

Une fois calculer, il suffit de remplacer toutes les valeurs manquantes par la moyenne, dans le cas ou l'une de nos attributs possède des valeurs catégoriques comme c'est le cas pour EnvironmentSatisfaction, nous pouvons soit opter pour une approche pessimiste (en prenant le minimum) ou optimiste (en prenant le maximum).
La seconde mesure qui peut être utilisé est la médiane dont on rappelle la formule vu dans la partie précédente :

$$Mediane = L1 + \frac{\frac{n}{2} + (\sum f)^l}{fMediane}$$

Encore une fois calculé, il suffit de la mettre à la place des valeurs manquantes.
Pour finir, la dernière mesure étant le mode qui la valeur la plus fréquente au sein de notre attribut. Dans le cas où nous avons plusieurs modes, n'importe lequel peut être utilisé.

1.1.2 En utilisant la substitution

Deux méthodes distinctes sont disponibles pour ça.
Le Forward filling : qui consiste à remplacer la valeur manquante par la valeur existante qui la précède.
Le Backward filling : qui consiste à remplacer la valeur manquante par la valeur existante qui la succède.
Ces méthodes sont très appréciées pour leurs simplicités, cependant d'autres méthodes plus complexes peuvent souvent donner des résultats plus satisfaisant comme nous allons le constater.

1.1.3 En utilisant la régression

La régression recouvre plusieurs méthodes d'analyse statistique permettant d'approcher une variable à partir d'autres qui lui sont corrélées. Dans notre cas, nous utiliserons deux types de régressions. la régression linéaire pour prédire les valeurs continues de l'attribut MonthlyIncome. Et la régression logistique pour prédire les valeurs catégoriques de l'attribut EnvironmentSatisfaction.

Régression linaire

régression linéaire est un modèle de régression qui cherche à établir une relation linéaire entre une variable, dite expliquée, et une ou plusieurs variables, dites explicatives.

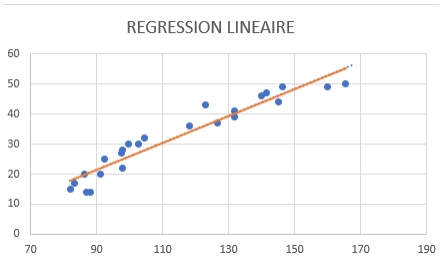


FIGURE 1.3 – Régression linéaire

Ce modèle nous permettra de faire de prédiction avec un taux de précision élevé nous permettant d'obtenir des nouvelles valeurs très convaincantes.

Régression logistique

La régression logistique est un modèle statistique permettant d'étudier les relations entre un ensemble de variables qualitatives Xi et une variable qualitative Y. Il s'agit d'un modèle linéaire généralisé utilisant une fonction logistique comme fonction de lien.

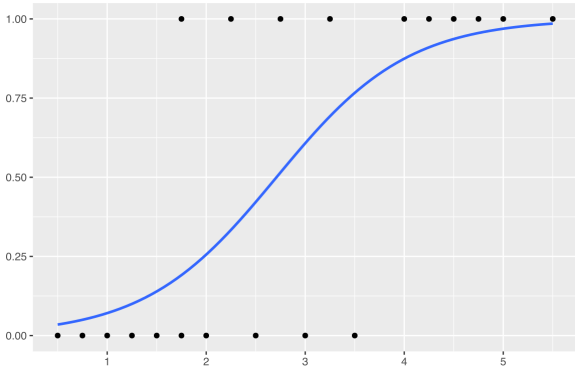


FIGURE 1.4 – Régression logistique

Tout comme la régression linéaire, cela nous permettra de remplacer nos valeurs manquantes par des prédictions avec un haut taux de précision.
Comme on peut le constater, même si ces méthodes sont plus complexes, elles permettent d'obtenir des résultats plus convaincants et moins redondant que les autres méthodes plus simples à implémenter.

1.2 Traitement des valeurs aberrantes

Les valeurs aberrantes sont des valeurs distantes comparées à la majorité des valeurs d'un attribut. Ces valeurs peuvent s'avérer problématiques dans de nombreux cas et il est important de

les prendre en charge afin d'éviter toutes erreurs ou anomalies dans de futur traitement. Dans le cas où elles sont peu nombreuses. On peut se permettre de juste les supprimer. Mais dans la majorité des cas, il est préférable de les remplacer par l'une des mesures de tendances. Par exemple, pour l'attribut YearsWithCurrManager une fois les valeurs aberrantes remplacé par la moyenne, nous obtenons la boite à moustache suivante :

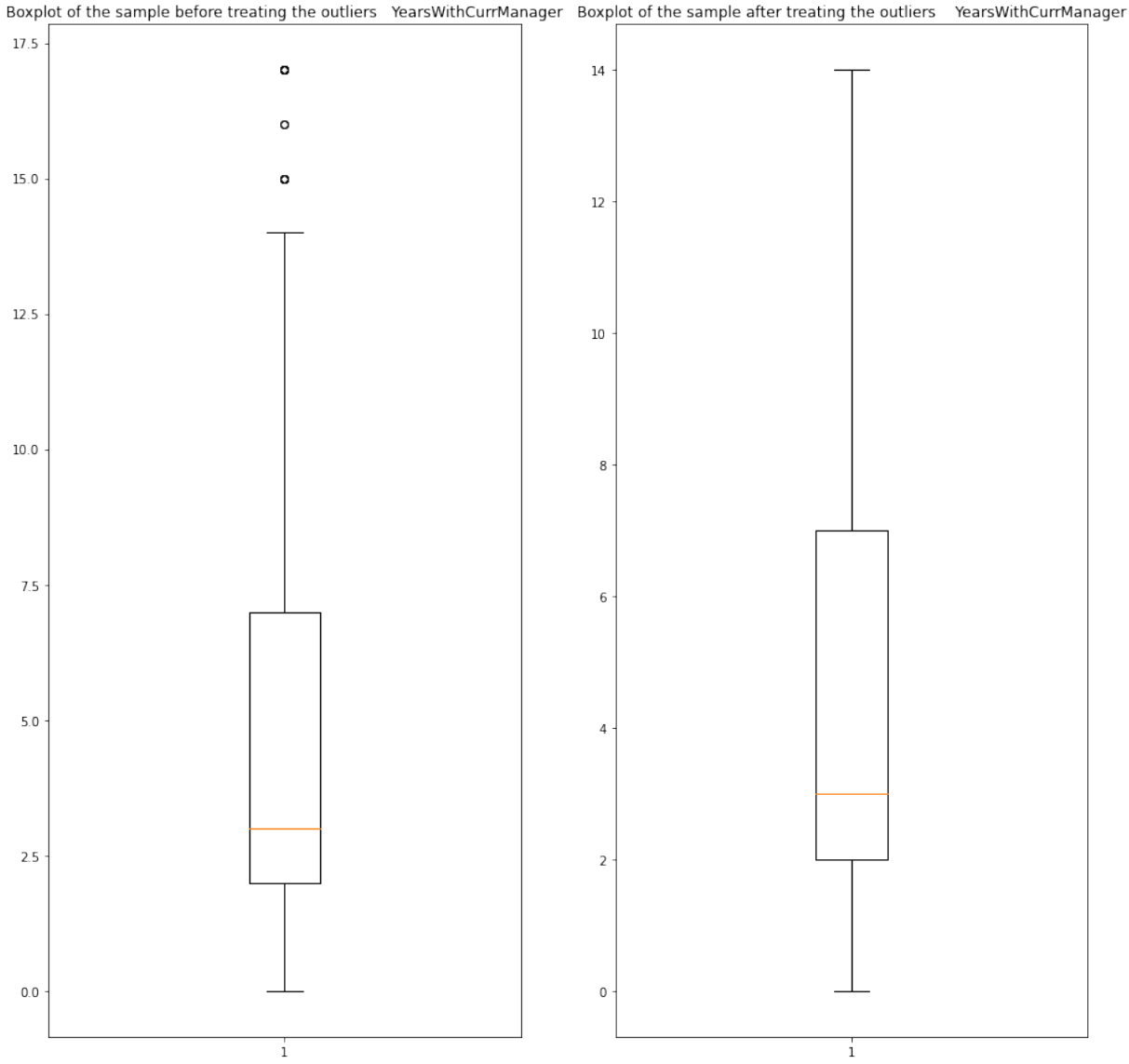


FIGURE 1.5 – Boite à moustache de YearsWithCurrManager sans et avec outliers

Comme on peut le constater, l'ensemble des données aberrantes à disparu.

Réduction des données

Introduction Ce chapitre nous introduit à la réduction des données qui consiste à transformer nos données brutes en des données utilisables pour différents traitements et analyses. Nous nous attarderons sur différente manière de faire disponible.

2.1 Réduction des données via la discrétisation

La discrétisation est une méthode permettant de transformer les données continues en un nombre fini de classes. dans ce chapitre, nous nous attarderons sur deux méthodes en particulier.

2.1.1 Discrétisation en classes d'amplitudes égales

Utilisée pour les formes de distribution uniforme et symétrique. Cela consiste à avoir un écart égale entre toutes les valeurs extrêmes (minimum et maximum d'une classe). Cette méthode possède de nombreux atouts, tels qu'elle permet une meilleure répartition des valeurs, peut gérer les données aberrantes et peut être combiné avec des données catégoriques.

Une fois appliquée, voici le résultat sur quelques attributs continue de note dataset.

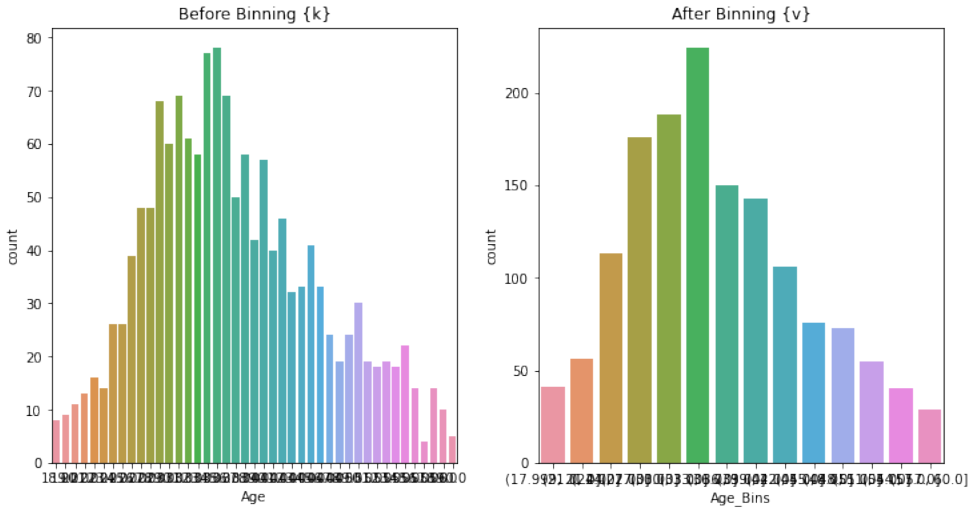


FIGURE 2.1 – Discrétisation par amplitude égale de l'attribut Age

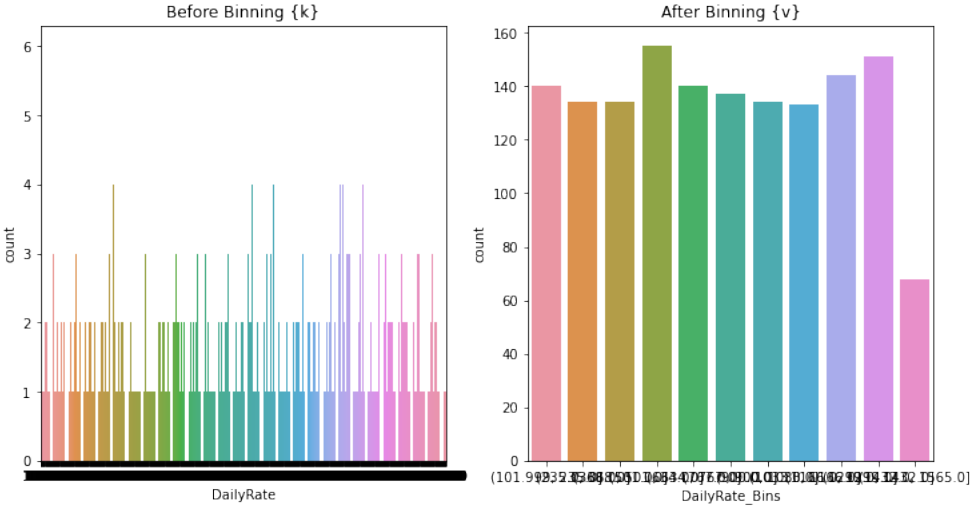


FIGURE 2.2 – Discrétisation par amplitude égale de l'attribut DailyRate

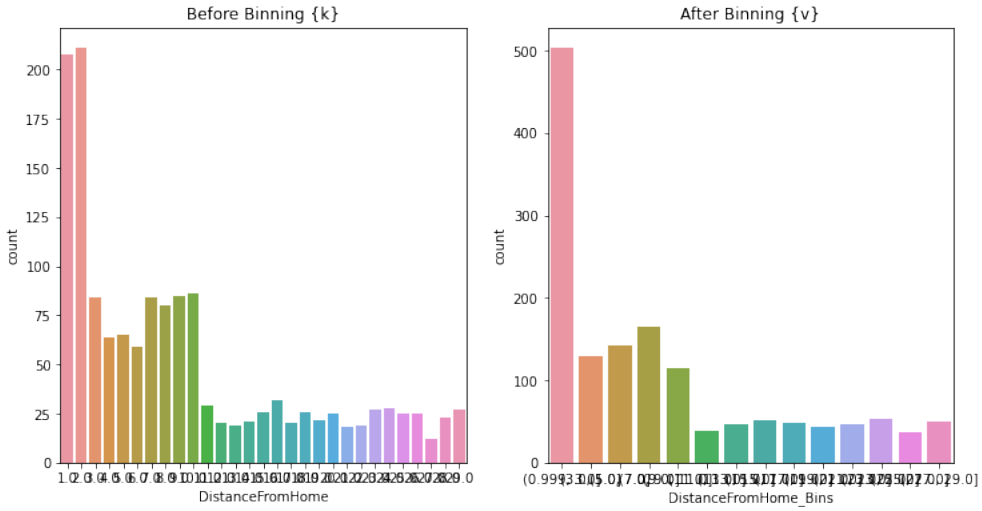


FIGURE 2.3 – Discrétisation par amplitude égale de l'attribut DistanceFromeHome

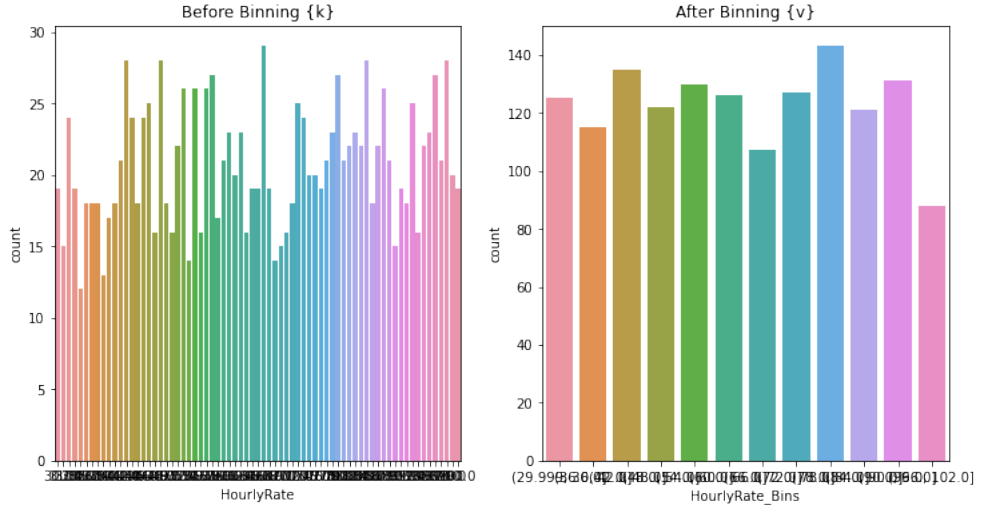


FIGURE 2.4 – Discrétisation par amplitude égale de l'attribut HourlyRate

Comme on peut le constater, cela ne donne pas des classes égales au niveau des valeurs appartenant à la classe, ce qui offre une distribution variée qui peut être utile dans de nombreux cas.

2.1.2 Discrétisation en classes d'effectifs égaux

Utilisée pour les formes de distribution dissymétrique et bimodale. Elle est caractérisée par une répartition équilibrée des individus. Avec l'effectif d'une classe $= N/k$ avec N le nombre totale d'individus et k le nombre de classes. Cette méthode est très appréciée pour sa répartition équilibrée des valeurs. Une fois appliquée, voici le résultat sur quelques attributs continue de note dataset.

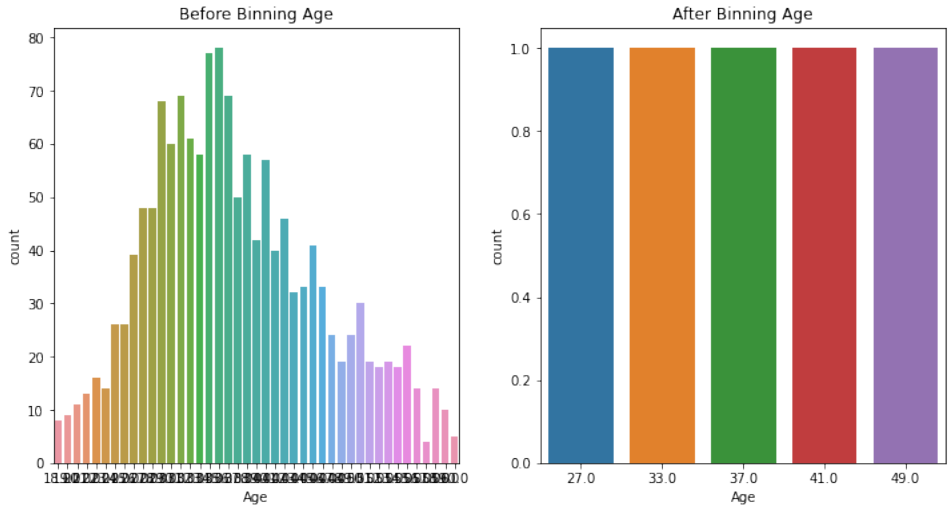


FIGURE 2.5 – Discrétisation par effectif égale de l'attribut Age

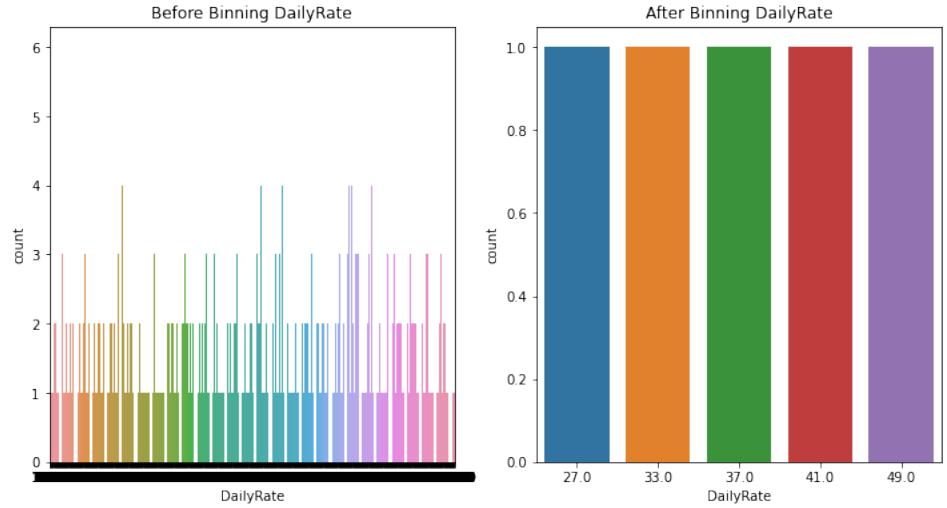


FIGURE 2.6 – Discrétisation par effectif égale de l'attribut DailyRate

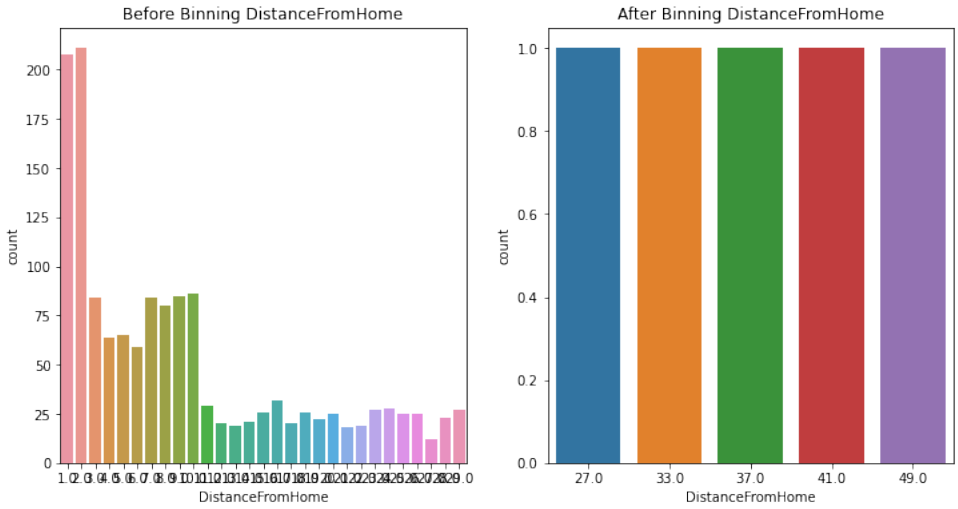


FIGURE 2.7 – Discrétisation par effectif égale de l'attribut DistanceFromeHome

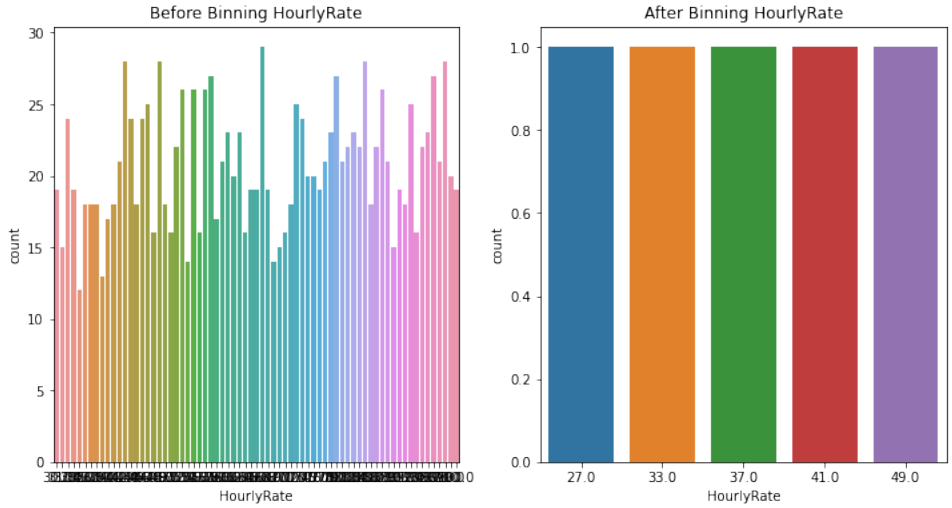


FIGURE 2.8 – Discrétisation par effectif égale de l'attribut HourlyRate

Contrairement à la méthode précédente, on peut voir que chaque classe est réparti de manière égale et contiennent tous le même nombre d'individus. Chacune des discrétisations présentées a ses avantages et a un intérêt particulier selon l'attribut et nos besoins.

2.2 Réduction des données via l'élimination des redondances

L'élimination des données redondantes consiste à supprimer des données doublons ou des attributs trop similaires ou peu importantes et ce processus se sépare en deux partis.

2.2.1 Élimination des redondances horizontale

Cette méthode consiste simplement à supprimer les lignes identiques de notre dataset. N'ayant pas de cas du genre, cette méthode n'apporte aucun changement.

2.2.2 Élimination des redondances verticale

La première étape consiste à supprimer les attributs n'ayant qu'une seule valeur ayant été détectée lors de la partie 1. Par la suite, nous utilisons des techniques de feature selection afin de conserver les attributs les plus pertinents. Dans notre cas, nous optons pour les techniques suivantes :

2.2.3 Méthode chi2

Cette méthode permet d'apporter une solution au problème de feature sélection en testant les relations entre les différents attributs.
La première étape consiste à créer la table de contingence (est une méthode de représentation de données permettant d'estimer la dépendance entre deux) entre nos attributs.
La deuxième étape, elle, consiste à calculer le degré de liberté qui représente le nombre d'observations moins le nombre de relations entre ces observations.
La troisième étape revient à calculer la valeur attendue de chaque valeur qui est égale à :

$$E(i,j) = (table_contingence[i][j] * somme_colonne)/somme_total$$

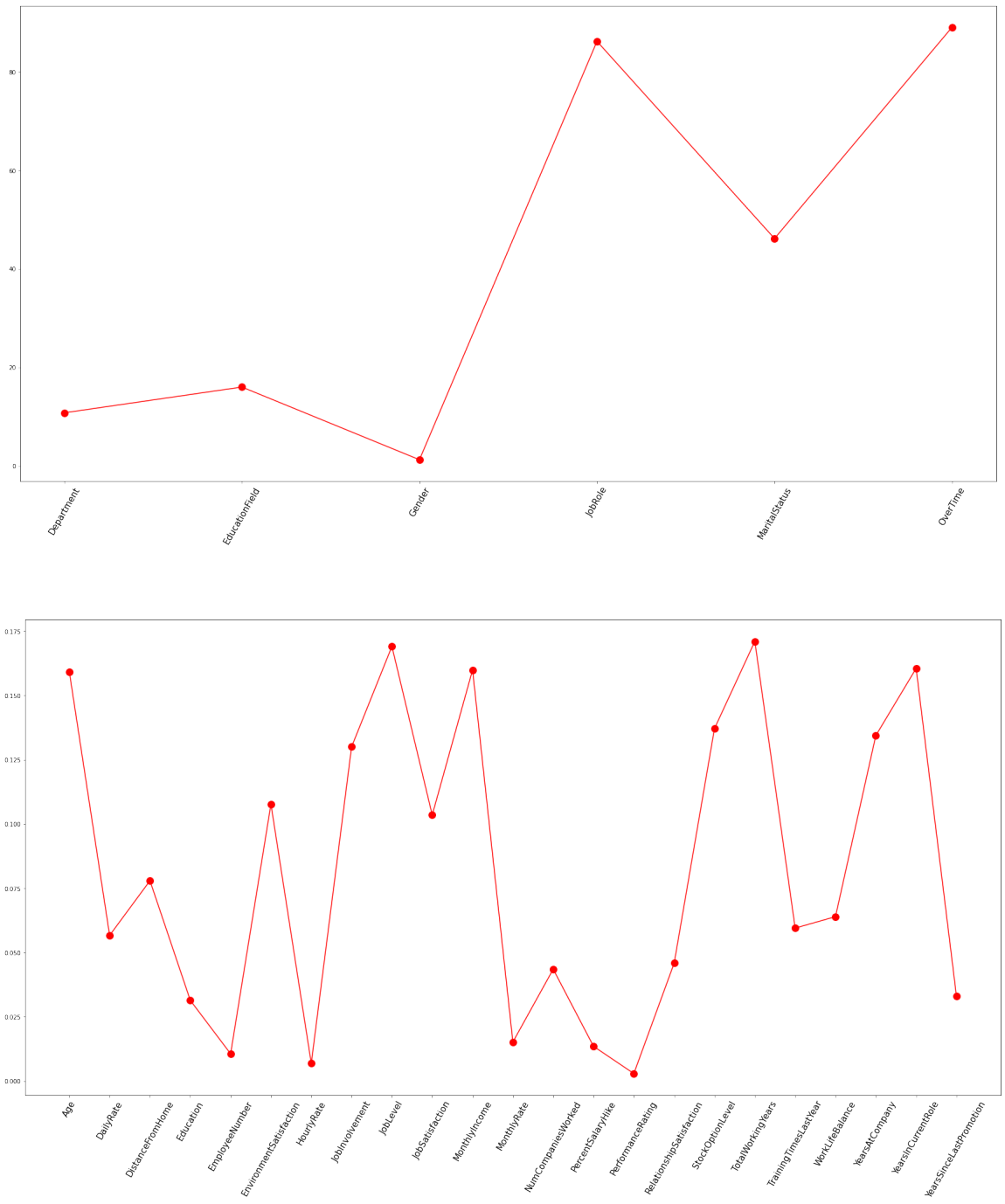
l'étape suivante consiste à calculer chi-square qui se calcule avec la formule suivante :

$$chi - square = \sum_{i=1}^{n} (obeserved - E_i)^2/E_i$$

Une fois fait, il suffit de calculer le chi-square score et estimer si la valeur est satisfaisante ou non.

2.2.4 Méthode Point-biserial

Cette méthode permet de calculer le coefficient de corrélation entre un attribut et un autre, par exemple en utilisant l'attribut Attrition avec les autres attributs, nous obtenons les corrélations suivantes :



Ces 2 méthodes présentées nous permettent d'obtenir les attributs les plus pertinents à garder.

Normalisation de données

Introduction Normalisation : La normalisation est une méthode de prétraitement des données qui permet de réduire la complexité des modèles. C'est également un préalable à l'application de certains algorithmes. Dans ce chapitre, nous nous attarderons sur deux méthodes de discrétisation.

3.1 Méthode MinMax

La méthode minmax consiste en la formule suivante :

$$Val_normal = (Val - min)/(max - min)$$

Avec :
Val-normal : La valeur normalisée
Val : La valeur avant normalisation
min = La valeur minimum de la colonne
max = La valeur maximum de la colonne
Une fois appliqué, nous obtenons les résultats suivants :

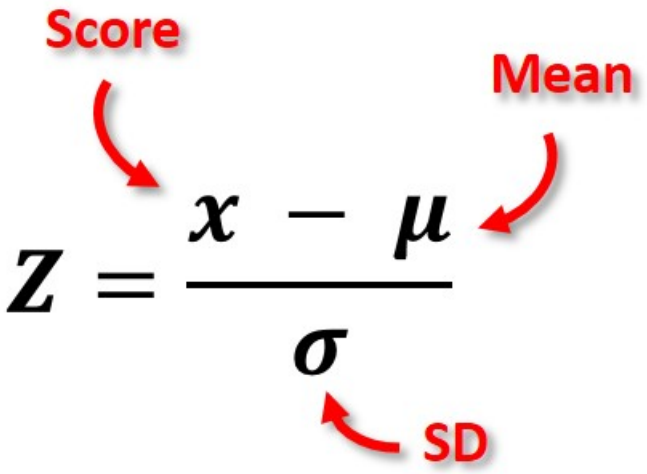
Age	DailyRate	DistanceFromHome	Education	EmployeeNumber	EnvironmentSatisfaction	HourlyRate	JobInvolvement
0.547619	0.715820	0.000000	0.25	0.000000	0.333333	0.914286	0.666667
0.738095	0.126700	0.250000	0.00	0.000484	0.666667	0.442857	0.333333
0.452381	0.909807	0.035714	0.25	0.001451	1.000000	0.885714	0.333333
0.357143	0.923407	0.071429	0.75	0.001935	1.000000	0.371429	0.666667
0.214286	0.350036	0.035714	0.00	0.002903	0.000000	0.142857	0.666667

FIGURE 3.1 – Echantillon de la méthode MinMax

Nous pouvons constater que toutes les valeurs se retrouvent a la même échelle.

3.2 Méthode Zscore

Cette méthode de normalisation utilise la formule suivante :



Une fois appliqué, nous obtenons les résultats suivants :

Age	DailyRate	DistanceFromHome	Education	EmployeeNumber	EnvironmentSatisfaction	HourlyRate	JobInvolvement
0.446199	0.742274	-1.010565	-0.891385	-1.700704	-0.667057	1.382668	0.379543
1.321915	-1.297333	-0.147100	-1.867790	-1.699043	0.249367	-0.240595	-1.025818
0.008340	1.413882	-0.887213	-0.891385	-1.695721	1.165791	1.284288	-1.025818
-0.429518	1.460969	-0.763861	1.061426	-1.694060	1.165791	-0.486544	0.379543
-1.086306	-0.524116	-0.887213	-1.867790	-1.690738	-1.583481	-1.273580	0.379543

FIGURE 3.2 – Echantillon de la méthode Zscore

Contrairement à la méthode précédente, nous pouvons constater des valeurs négatives et supérieures a 1. l'utilisation de l'une de ses 2 méthodes dépend des besoins.

Conclusion Générale

Dans cette partie du projet, nous nous sommes introduits au prétraitement des données, ce qui nous a permis de régler les intégralités des anomalies trouvées lors de la partie 1, mais aussi de préparer le dataset a de futur traitement qui prendront tous leurs sens lors des parties suivantes.