

Module : Statistique pour Machine Learning

Travaux Pratiques : Séance 1

Objectifs à atteindre :

- Etre en mesure de créer une banque de données dans l'environnement SPSS par la déclaration des variables (étiquettes de variable, type, étiquette de valeur, format et valeurs manquantes).
- Etre en mesure de corriger les données et d'introduire de nouvelles variables.

Cette première séance de travail consiste à utiliser SPSS afin de préparer une banque de données provenant d'une enquête réalisée auprès d'une catégorie d'employés d'une agence d'une banque donnée. Vous trouverez dans la figure 1, un aperçu de ces données. Le dictionnaire de données et quelques données n'ont pas été créés, vous devez donc le faire vous-même directement dans l'environnement de SPSS.

ID	SALDEB	SEXE	TEMPS	AGE	SALACT	NIVETUD	CATEMP	STATUT
1	7200	m	79	46,58	11460	15	1	0
2	6900	M	79	28,00	16080	15	1	0
3	5400	m	67	28,75	14100	15	1	0
4	5040		96	27,42	12420	15	1	0
5	6300	m	84	33,50	15720	15	1	7
6	600	m	88	54,33	8880	12	1	0
7	6900	f	72	32,67	10380	15	1	1
8	6300	f	70	58,50	8520	15	1	0
9	6000	f	84	,00	15540	15	1	0
10	5400	m	97	31,92	10920	12	1	1
11	7800	m	83	38,42	11736	12	1	1
12	6600	m	95	33,75	14040	15	1	0
13	7500	m	98	35,67	16080	16	1	0
14	7800	f	78	28,33	16140	16	1	1
15	8160	F	84	295,00	15000	15	1	0

Tableau 1.1 : Description des données de la base

Description de la codification de la banque de données

Champs	Etiquette	Valeurs
ID	Code de l'employé(e)	
SALDEB	Salaire d'embauche	
SEXE	Sexe de l'individu	m : Homme, f :Femme
TEMPS	Ancienneté employé	
AGE	Age de l'employé(e)	
SALACT	Salaire actuel	
NIVETUD	Niveau d'études	
CATEMP	Catégorie d'employé(e)	1: Employé de bureau, 2: Employé stagiaire, 3: Agent de sécurité, 4: Rédacteur stagiaire, 5: Personnel vacataire, 6: Cadre stagiaire, 7: Personnel technique
STATUT	Souscripteur	0 : Oui, 1 : Non

Tableau 1.2 : Description des données de la base

Voici les étapes à suivre pour réaliser dans cette séance:

1- Création du dictionnaire de données

Démarrer SPSS et définissez la structure de données (nom des variables ainsi que leurs types, textes explicatifs et valeurs manquantes) telle qu'elle est décrite plus haut. N'oubliez pas d'enregistrer ensuite cette structure de fichier dans le fichier agence.sav. Dans la boîte de dialogue « *Ouvrir un fichier* »

A noter qu'il est recommandé de prendre connaissance des variables qui constituent le fichier en affichant le dictionnaire de données de la manière suivante :

Dans le menu Fichier, cliquez sur **Afficher des informations sur les fichiers de données> Fichier de travail.**

Vous devrez maintenant compléter le dictionnaire de données par la déclaration des étiquettes de variables tel qu'il est spécifié dans le tableau 1.1. Pour créer le dictionnaire de données, cliquez sur l'onglet « **Affichage des variables** » en bas de

Module : Statistique pour Machine Learning

la fenêtre de données. Continuez à compléter le dictionnaire de données des autres variables en spécifiant leurs étiquettes ainsi que leur étiquettes de valeurs si elles existent.


Une fois que vous avez fini de compléter le dictionnaire de données, générer encore une fois le récapitulatif du dictionnaire de données, afin de vérifier les informations saisies.

2- Correction des données

Dans cet exercice vous devez vérifier la fiabilité des données de l'agence et traiter les anomalies détectées.

Créez un tableau d'effectif de la variable sexe en suivant les étapes suivantes :

Dans le menu **Analyse**, allez dans **Statistiques descriptives** puis cliquez sur **Fréquences**

Sélectionnez la variable « **Sexe de l'individu** » puis cliquez sur le bouton  pour la faire passer dans la colonne '**Variable(s)**'. Cliquez sur le bouton **OK**.

Vous remarquez que SPSS différencie les valeurs alphanumériques qui sont majuscules ou minuscules. Vous remarquez aussi que les valeurs alphanumériques vides sont considérées comme des valeurs valides et non pas comme des valeurs manquantes par défaut.

Pour remédier à ces anomalies, vous devez spécifier la valeur vide en tant que valeur manquante dans le dictionnaire de données de la variable Sexe.

Allez dans la fenêtre **Affichage des variables**

Cliquez dans la case **Manquant** de la variable **Sexe**

Sélectionnez le choix **Valeurs manquantes discrètes**


Tapez sur la **barre d'espace** dans le premier champ

Puis cliquez sur le bouton **OK**.

Produisez maintenant une autre table de fréquence (ou effectifs) sur la variable Sexe, vous constatez que le champ vide est maintenant considéré comme donnée manquante.

Il suffira maintenant de standardiser F et M en f et m.

Dans le menu **Transformer**, sélectionnez **Recoder>Recodage de Variables...**

Dans la liste des variables, sélectionnez la variable Sexe de l'individu et cliquez sur le bouton 

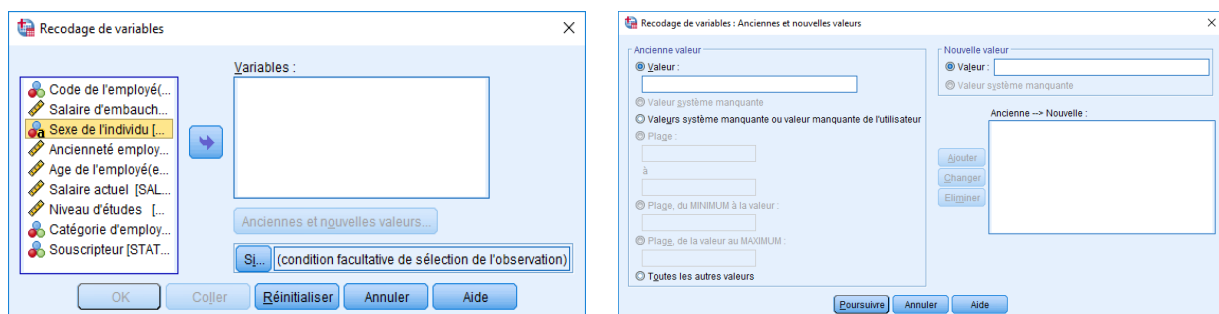
Cliquez sur le bouton **Anciennes et nouvelles valeurs...**

Dans le champ Valeur de la partie **Ancienne valeur**, saisissez la valeur **F**

Puis dans le champ Valeur de la partie **Nouvelle valeur**, saisissez la valeur **f**

Cliquez sur le bouton **Ajouter**.

Faites maintenant de même pour la valeur **M** en la recodant en **m**.




Boîte de dialogue Recodage de variables.

Une fois que vous avez fini le recodage, cliquez sur le bouton **Poursuivre**, puis sur le bouton **OK**.

Maintenant générez un autre tableau d'effectifs pour la variable **sexe**, afin de vérifier que le recodage a bien été exécuté.

Module : Statistique pour Machine Learning

Dans cet exercice vous avez traité une variable alphanumérique en recodant les valeurs et spécifiant la valeur manquante. La suite de cet exercice vous permettra de connaître les techniques de traitement des variables numériques.

Dans le menu **Analyse**, allez dans **Statistiques descriptives>Descriptives...**
Sélectionnez la variable **Age de l'employé** puis cliquez sur le bouton 
Cliquez sur le bouton **OK**.

Vous notez que la valeur minimale de l'âge est **0** et la valeur maximale est de **295**. Ces valeurs sont évidemment erronées, et vous allez faire les traitements suivants pour corriger ces anomalies.

Concernant la valeur nulle, elle détermine une valeur non renseignée de l'âge.
Allez dans l'**Affichage des variables**, puis cliquez sur le champ **Manquant** de la variable **Age**
Sélectionnez **Valeurs manquantes discrètes**, et saisissez la valeur **0**
Cliquez sur le bouton **OK**.

Pour corriger l'âge de l'employé avec l'ID numéro 15, vous allez utiliser la fonction Aller à l'observation de SPSS. Cette technique est faisable lorsque la ligne où se trouve l'anomalie est connue. Dans le cas où plusieurs cas atypiques existent dans les données, vous devez alors entreprendre des analyses des valeurs atypiques pour chaque variable. D'autres modules de formations SPSS traitent la manipulation avancée des données.

Dans la fenêtre **Affichage des données**, cliquez sur la colonne **Age**
Allez dans le menu **Edition**, et sélectionnez **Aller à l'observation...**
Dans le champ **Numéro de l'observation**, tapez la valeur 15
Cliquez sur le bouton **OK**.

Vous êtes maintenant dans la ligne numéro 15, dans la position où se trouve la valeur atypique.

Dans la case **Age**, remplacez la valeur **295** par la valeur **29.5**.
Produisez un nouveau tableau descriptif pour vérifier la fiabilité de la variable Age.
Dans la variable **Statut**, remplacez par recodage, la valeur 7 par la valeur 1.

3- Création d'une variable et saisie de données

ID	EXP
1	21.75
2	3.17
3	0.50
4	1.17
5	6.00
6	27.00
7	6.92
8	31.00
9	4.42
10	5.50
11	12.50
12	7.50
13	3.00
14	1.67
15	2.75

Dans cet exercice, vous allez pouvoir créer une nouvelle variable nommée EXP qui indique la durée d'expérience de la personne dans la vie professionnelle. Cette variable doit être insérée entre les variables NIVETUD et CATEMP.

Voici les données relatives à la durée d'expérience de chaque employé.

Pour insérer la variable **EXP**, sélectionnez la colonne **CATEMP**

Allez dans le menu **Edition** et sélectionnez **Insérer une variable**

Une nouvelle colonne apparaît maintenant entre la colonne NIVETUD et CATEMP. Spécifiez, dans la fenêtre Affichage de variable, le dictionnaire de données de cette nouvelle variable en lui indiquant les spécifications suivantes :

Nom : EXP

Etiquette : Expérience professionnelle (an)

Ensuite dans la fenêtre Affichage des données, saisissez par ordres les valeurs de la variable EXP pour chaque employé en respectant le numéro d'identification.

Une fois la saisie est complète, générez la description du dictionnaire de données
Pour enregistrer les changements effectués sur le fichier de travail,

Allez dans le menu **Fichier**, puis cliquez sur **Enregistrer sous...**

Dans le champ **Nom du Fichier**, saisissez **AgenceTP.sav**

Module : Statistique pour Machine Learning

Cliquez sur le bouton **OK** pour validation.

4- Utilisation du modèle de données

La création du dictionnaire de données dans un fichier de données SPSS permet à l'utilisateur d'utiliser un fichier de données SPSS externe comme modèle pour définir les propriétés de fichier et de variable dans le fichier de travail courant.

Ouvrir le fichier **Agence_total.sav**

Générer le dictionnaire de données du fichier de travail

Vous remarquez que le dictionnaire de données est vide et les variables sont similaires à celles utilisées dans le fichier précédent. Vous allez maintenant appliquer le dictionnaire de données créé dans le fichier précédent dans le fichier de travail actuel.

Dans le menu Données, sélectionnez **Copier les propriétés de données...**

Dans le choix de la source des propriétés, sélectionnez l'option **Un fichier de données SPSS externe**

Cliquez sur le bouton **Parcourir** pour spécifier le chemin du fichier de données SPSS **agence_TP.sav**

Dans la deuxième étape de la copie du dictionnaire de données, sélectionnez le premier choix, qui permet d'appliquer les propriétés des variables sélectionnées dans le fichier source aux variables concordantes dans le fichier de travail.

Vérifier que toutes les variables sélectionnées du fichier source ont un correspondant dans le fichier de travail - **Variables concordantes** : 10

Cliquez trois fois sur le bouton **suivant**, jusqu'à l'étape 5, puis cliquez sur le bouton **Fermer**.

Vous remarquez que les propriétés des variables ont reçu celles du fichier de données **SPSS agence_tp.sav**.

Enregistrer le fichier actuel sous le nom de : **Agence_TP_Total.sav**