Rapport concernant les données - 09/11/2023

Projet tutoré

Océane Baboulaz, Alice Chauveau
Charline Champ et Morgane Maulet

M2 SSD
Année universitaire 2023/2024

Contents

Introduction

Ce fichier a pour but d'expliquer au mieux ce que nous avons commencé à traiter sur le jeu de données qui nous a été donné. Tout le code est disponible dans notre fichier *"Data/mise_en_formeV2.R"* que vous pouvez retrouver sur notre git.

Ainsi nous allons vous expliquer les étapes que nous avons réalisée :

- Gestion des numéros de dossiers vides
- Gestion des doublons
- Modification sur les colonnes
- Regroupement de données et calcul des scores
- Gestion des types de variables

A noter que sans aucunes modifications notre jeu de données contient 431 coureurs pour 145 variables.

Gestion des numéros de dossiers vides

Lors de notre première réunion avec Clément Baud, nous avons remarqué que nous avions des noms de dossiers vides contenant des informations. Parmi cela, nous avions des lignes vides que nous devions supprimer. Pour mieux visualiser voici le tableau des 23 cas où nous avons une valeur “NA” pour notre numéro de dossier. Remarquons que nous visualisons ici que les 5 premières colonnes.

Table 1: Visualisation d’un bout des données contenant des numéros de dossiers vides

Q64	Q1	Q2	Q3.1	Q3.2
NA	Femme	46	54	54
NA	Homme	46	63	64
NA	Homme	63	80	82
NA	Homme	55	75	77
NA	Homme	63	64	65
NA	Femme	25	58	59
NA	NA	NA	NA	NA
NA	NA	NA	NA	NA
NA	Homme	34	80	85
NA	NA	NA	NA	NA
NA	NA	NA	NA	NA
NA	Homme	49	64	65
NA	Femme	36	54	56
NA	Homme	43	77	77
NA	Homme	55	83	85
NA	NA	NA	NA	NA
NA	NA	NA	NA	NA
NA	NA	NA	NA	NA
NA	Homme	39	66	70
NA	NA	NA	NA	NA
NA	NA	NA	NA	NA
NA	NA	NA	NA	NA
NA	NA	NA	NA	NA

A l’aide de ce tableau, nous pouvons observer que parmi les 23 valeurs “NA” nous obtenons 11 lignes qui ne contiennent rien. Nous les avons donc supprimées. Parmi les 12 restantes, et donc les 12 qui contiennent de l’information, nous avons décidé de leur attribuer un numéro de dossier écrit de la manière suivante : “NA_” + le numéro de sa ligne.
Vous trouverez ci-dessous le même tableau que vu précédemment avec les modifications effectuées.

Table 2: Visualisation d'un bout des données après modification des numéros de dossiers vides

Q64	Q1	Q2	Q3.1	Q3.2
NA_60	Femme	46	54	54
NA_95	Homme	46	63	64
NA_99	Homme	63	80	82
NA_111	Homme	55	75	77
NA_181	Homme	63	64	65
NA_203	Femme	25	58	59
NA_223	Homme	34	80	85
NA_227	Homme	49	64	65
NA_256	Femme	36	54	56
NA_289	Homme	43	77	77
NA_300	Homme	55	83	85
NA_400	Homme	39	66	70

A noter que nous avions bien pensé, avant de les renommer, de vérifier que ce n'était pas des doublons.

Gestion des doublons

Une fois la première étape réalisée, nous nous sommes attelées à la vérification de doublons dans nos données. En effet, lorsque nous avons regardé nos numéros de dossier nous nous sommes rendues compte que des numéros de dossiers avaient remplis deux, trois voire même quatre fois le questionnaire. Vous trouverez ci-dessous les 85 doublons que nous avons trouvé.

Table 3: Visualisation des doublons dans notre jeu de données

Q64	Q1	Q2	Q3.1	Q3.2
AHF30789798	Homme	33	69	70
AHF30789798	Homme	33	69	71
BJP48436333	Homme	25	78	80
BJP48436333	Homme	25	79	82
BTK83111539	Homme	47	74	75
BTK83111539	Homme	48	74	75
CCI09116081	Femme	23	59	61
CCI09116081	Femme	23	58	60
CTW25591815	Homme	22	72	75
CTW25591815	Homme	22	70	75
CTW25591815	Homme	22	70	75
DHP89151690	Homme	32	75	76
DHP89151690	Homme	33	75	76
DQL60938665	Homme	23	73	78
DQL60938665	Homme	23	73	77
DQL60938665	Homme	23	73	78
EVZ71437549	Homme	34	79	85
EVZ71437549	Homme	34	78	83
FED95806025	Homme	66	76	79
FED95806025	Homme	66	76	79
FWU27646825	Homme	57	72	74
FWU27646825	Homme	57	72	74
GAS08674864	Homme	40	68	69
GAS08674864	Homme	40	68	69
GAS08674864	Homme	40	68	70
GAS08674864	Homme	40	68	70
IMJ55075863	Homme	47	71	74
IMJ55075863	Homme	57	71	74
JFS53974621	Femme	35	55	60
JFS53974621	Femme	35	55	60
JQB98921250	Femme	25	53	53
JQB98921250	Femme	25	53	53
JXG10263729	Homme	43	78	90
JXG10263729	Homme	44	80	85
LHK28906512	Homme	32	85	88
LHK28906512	Homme	31	85	89
LJW27611575	Homme	48	82	82
LJW27611575	Homme	48	82	82
LOD42209851	Femme	64	58	59
LOD42209851	Femme	64	58	59
MDB15723690	Homme	54	66	68

Q64	Q1	Q2	Q3.1	Q3.2
MDB15723690	Homme	54	66	68
MYZ21599350	Homme	41	63	66
MYZ21599350	Homme	41	63	66
NII32600142	Femme	25	53	56
NII32600142	Femme	25	53	56
NSL16760077	Homme	51	94	97
NSL16760077	Homme	51	94	97
OWG60593535	Homme	21	68	68
OWG60593535	Homme	21	65	68
QHY23545742	Homme	34	72	72
QHY23545742	Homme	34	72	72
QIQ45303215	Homme	33	73	75
QIQ45303215	Homme	33	72	75
QLR76211017	Homme	35	82	83
QLR76211017	Homme	35	81	83
REE83272089	Homme	65	72.2	75
REE83272089	Femme	65	72.2	75
RHJ73628227	Femme	52	61	62
RHJ73628227	Femme	52	59	62
SDM17591012	Homme	33	62	62
SDM17591012	Homme	33	62	62
SVP60626908	Homme	26	78	83
SVP60626908	Homme	26	78	83
TNJ60561654	Homme	28	64	64
TNJ60561654	Homme	28	64	64
UBO49849223	Homme	51	68	68
UBO49849223	Homme	52	68	68
UGT86942227	Homme	49	72	76
UGT86942227	Homme	49	73	77
UUU26690739	Homme	42	80	86
UUU26690739	Homme	43	80	83
VHA06723427	Homme	34	77	76
VHA06723427	Homme	34	77	76
VLX28454772	Homme	24	75	76
VLX28454772	Homme	24	75	76
XKI10591470	Femme	42	57	57
XKI10591470	Femme	42	57	57
XKI10591470	Femme	42	57	57
YDN52180692	Femme	41	54	58
YDN52180692	Femme	41	54	58
YUL56925637	Femme	37	63	66
YUL56925637	Femme	37	64	66
ZMS73503351	Homme	47	76	78
ZMS73503351	Homme	47	75	80

Nous avons ensuite représenté nos doublons en fonction de la variable Finished. Pour rappel, cette variable vaut “True” si le coureur a fini de remplir le questionnaire et “False” si ce n’est pas le cas. Le tableau ci-dessous vous illustre cela.

Table 4: Doublons en fonction de la variable Finished

	False	True
AHF30789798	0	2
BJP48436333	1	1
BTK83111539	1	1
CCI09116081	1	1
CTW25591815	1	2
DHP89151690	1	1
DQL60938665	2	1
EVZ71437549	0	2
FED95806025	1	1
FWU27646825	0	2
GAS08674864	2	2
IMJ55075863	1	1
JFS53974621	2	0
JQB98921250	1	1
JXG10263729	0	2
LHK28906512	1	1
LJW27611575	1	1
LOD42209851	1	1
MDB15723690	1	1
MYZ21599350	1	1
NII32600142	1	1
NSL16760077	1	1
OWG60593535	1	1
QHY23545742	1	1
QIQ45303215	0	2
QLR76211017	0	2
REE83272089	1	1
RHJ73628227	1	1
SDM17591012	1	1
SVP60626908	1	1
TNJ60561654	1	1
UBO49849223	1	1
UGT86942227	1	1
UUU26690739	0	2
VHA06723427	1	1
VLX28454772	1	1
XKI10591470	3	0
YDN52180692	1	1
YUL56925637	0	2
ZMS73503351	1	1

Remarquons que nous obtenons 3 configurations différentes :

- Ceux qui ont des questionnaires remplis et non-remplis.
- Ceux qui n'ont que des questionnaires non-remplis.
- Ceux qui n'ont que des questionnaires remplis.

Etape 1 : Finished == "True" et Finished == "False"

Notre première étape a été de retirer les doublons qui avaient :

- Fait au moins une fois le questionnaire et remplit jusqu'au bout, ce qui implique une valeur "True" dans la variable Finished
- Fait une ou plusieurs autre fois le questionnaire mais cette fois-ci pas remplit jusqu'au bout, ce qui implique une valeur "False" dans la variable Finished.

Ainsi nous avons supprimé tous les cas qui avaient à la fois une observation pour un questionnaire remplit et une ou plusieurs pour un questionnaire non remplis en sélectionnant parmi ces cas uniquement les observations qui avaient remplit jusqu'au bout le questionnaire.
Attention, ici on parle d'observation et non pas d'individu car nous avons plusieurs données pour les individus.

Après sélection nous obtenons donc 53 cas et donc nous avons supprimé 32 cas. Voici le tableau précédemment vu avec ces modifications.

	False	True
AHF30789798	0	2
BJP48436333	0	1
BTK83111539	0	1
CCI09116081	0	1
CTW25591815	0	2
DHP89151690	0	1
DQL60938665	0	1
EVZ71437549	0	2
FED95806025	0	1
FWU27646825	0	2
GAS08674864	0	2
IMJ55075863	0	1
JFS53974621	2	0
JQB98921250	0	1
JXG10263729	0	2
LHK28906512	0	1
LJW27611575	0	1
LOD42209851	0	1
MDB15723690	0	1
MYZ21599350	0	1
NII32600142	0	1
NSL16760077	0	1
OWG60593535	0	1
QHY23545742	0	1
QIQ45303215	0	2
QLR76211017	0	2
REE83272089	0	1
RHJ73628227	0	1
SDM17591012	0	1
SVP60626908	0	1
TNJ60561654	0	1
UBO49849223	0	1
UGT86942227	0	1
UUU26690739	0	2
VHA06723427	0	1

	False	True
VLX28454772	0	1
XKI10591470	3	0
YDN52180692	0	1
YUL56925637	0	2
ZMS73503351	0	1

Maintenant que nous avons réalisé cette modification notre base de données contient 372 observations et il ne nous reste plus que 25 doublons à traiter. Voici ci-dessous notre nouveau tableau contenant les doublons toujours en fonction de la variable Finished.

Table 6: Doublons en fonction de la variable Finished après première modification

	False	True
AHF30789798	0	2
CTW25591815	0	2
EVZ71437549	0	2
FWU27646825	0	2
GAS08674864	0	2
JFS53974621	2	0
JXG10263729	0	2
QIQ45303215	0	2
QLR76211017	0	2
UUU26690739	0	2
XKI10591470	3	0
YUL56925637	0	2

Etape 2 : Finished == “False”

Une fois la première étape réalisée, nous avons pu nous occuper des doublons qui ne contenait plus que des valeurs égale à “False” dans la variable “Finished”. A l’aide du tableau suivant on remarque que nous avons 5 observations dans ce cas pour seulement 2 individus.

Pour pouvoir sélectionner lequel nous devions garder nous avons regardé lequel était le mieux remplit. Pour réaliser cela nous avons la variable Progress qui est une valeur entre 1 et 100 et nous renvoie le taux de progression dans le questionnaire que l’individu a réalisé.

Observons pour commencer le tableau des doublons concernant la valeur “False” pour la variable Finished en fonction de la variable Progress.

Progress	Q64	Q1	Q2	Q3.1	Q3.2
17	JFS53974621	Femme	35	55	60
88	JFS53974621	Femme	35	55	60
17	XKI10591470	Femme	42	57	57
17	XKI10591470	Femme	42	57	57
58	XKI10591470	Femme	42	57	57

Dans ce cas-là, nous avons donc gardé les lignes 2 et 5. Ainsi, à cette étape notre jeu de données contient 369 observations.

Etape 3 : Finished == "True"

Une fois avoir réalisé les deux premières étapes il nous reste à présent le cas où les doublons ont bien remplit les deux questionnaires jusqu'au bout. Nous obtenons 20 observations qui sont dans ce cas-là, et donc 10 coureurs.

Nous n'avons pas traiter ses cas là car nous devons voir avec vous comment nous pouvons faire.

Proposition : sélectionner les premiers questionnaires réalisés.

Une fois donc que cette étape sera réalisée, nous n'aurons plus de doublons dans notre jeu de données et nous aurons au final 359 coureurs.

Voici les identifiants des doublons concernaient.

Table 8: Doublons restants où nous devons prendre une décision

Q64	Q1	Q2	Q3.1	Q3.2
AHF30789798	Homme	33	69	70
AHF30789798	Homme	33	69	71
CTW25591815	Homme	22	72	75
CTW25591815	Homme	22	70	75
EVZ71437549	Homme	34	79	85
EVZ71437549	Homme	34	78	83
FWU27646825	Homme	57	72	74
FWU27646825	Homme	57	72	74
GAS08674864	Homme	40	68	69
GAS08674864	Homme	40	68	70
JXG10263729	Homme	43	78	90
JXG10263729	Homme	44	80	85
QIQ45303215	Homme	33	73	75
QIQ45303215	Homme	33	72	75
QLR76211017	Homme	35	82	83
QLR76211017	Homme	35	81	83
UUU26690739	Homme	42	80	86
UUU26690739	Homme	43	80	83
YUL56925637	Femme	37	63	66
YUL56925637	Femme	37	64	66

Modifications sur les colonnes

Suppression Nous avons supprimé les 10 premières colonnes qui ne nous sont pas utiles pour les analyses que nous allons réaliser par la suite.

Renommage Nous avons renommé les noms de colonnes pour que celles-ci ne contiennent ni d'accent ni d'espace. Cela sera beaucoup plus pratique ensuite pour nos analyses. Pour se faire, nous avons écrit nos noms de colonnes en anglais ce qui enlève le problème d'accent. Pour ce qui est du problème d'espace nous avons pris comme signe d'espace le symbole Under score : "_". Voici les 20 premières :

Table 9: Tableau des noms de colonnes avant et après modification

Noms de colonnes avant	Noms de colonnes après
Q64	file_number
Q1	sex
Q2	age
Q3.1	weight
Q3.2	max_weight
Q3.3	min_weight
Q70	size
Q4	weekly_volume_profession
Q5	start_trail
Q6	other_practice
Q66	other_practice_detail
Q7	time_proportion_other_practice
Q8	coach_for_trail
Q9.1	heart_rate_monitor_train
Q9.2	heart_rate_monitor_competition
Q10.1	connected_platform
Q11.1	ITRA_rating_know
Q11.2	ITRA_rating
Q12	number_hours_practice_trail_phase_preparation
Q13	drop_practice_trail_phase_preparation

Regroupement de données et calcul des scores

Pour tous les regroupements des questions et calcul des scores nous avons utilisé le dictionnaire des données que nous a transmis Clément Baud.
Pour se faire, nous avons créé une fonction nous permettant d'attribuer des numéros aux valeurs cochées pour le coureur.

Pour vous donner un exemple, si l'on regarde la question 35 concernant l'épanouissement qui est : Êtes-vous d'accord avec les propositions suivantes ? Vous avez 7 niveaux de réponses :

- Pas du tout d'accord
- Pas d'accord
- Plutôt pas d'accord
- Neutre
- Plutôt d'accord
- D'accord
- Tout à fait d'accord

Suivant la réponse donnée par sous réponse nous lui attribuons un chiffre entre 1 et 7. En disant que 1 est la modalité "Pas du tout d'accord" et 7 "Tout à fait d'accord". Une fois cela réalisé, nous avons des nombres et ainsi nous pouvons faire des moyennes. Ce processus a été réalisé pour plusieurs questions qui sont inscrites dans le fichier "Data/Dictionnaire_des_donnees".

Épanouissement

Méthodologie : Faire une moyenne, en utilisant les 8 réponses pour créer la variable « fulfillment »
Interprétation : Plus le score est élevé plus cela signifie que l'individu est épanouie

Priorités

Méthodologie : : Faire une moyenne en utilisant les résultats des 12 questions
Interprétation : Plus le score est élevé plus cela signifie qu'il existe des conflits entre vie personnelle et pratique de l'activité sportive

Motivation

Méthodologie : Il faut créer différents types de score de motivation (5) en moyennant de la manière suivante :
- Motivation intrinsèque : 2/7/12/17
- Motivation identifiée : 4/9/14/19
- Motivation intégrée : 3/8/13/18
- Motivation introjectée : 5/10/15/20
- Motivation externe : 6/11/16/21

A terme possibilité d'additionner certains types de motivation pour créer 2 variables : motivation intrinsèque et motivation extrinsèque

Interprétation : Suivant le type de motivation les raisons de pratique seront différentes, les comportements aussi

Force mentale

Méthodologie : Construction de différentes variables (4) - Force mentale générale : moyenne de la réponse à tous les items
- Confiance : moyenne des réponses aux items 13, 5, 11, 6, 14, 1

- Consistance : moyenne des réponses aux items : 3, 12, 8, 10
- Contrôle : moyenne des réponses aux items : 2, 4, 9, 7

Discours interne

Méthodologie : Construction de différentes variables
- Utilisation du Discours interne en général : moyenne de tous les items
- Utilisation du Discours interne à l'entrainement : moyenne des items 1, 2, 3, 4
- Utilisation du Discours interne en compétition : moyenne des items 5, 6, 7, 8

Interprétation : Plus le score est élevé plus l'individu à l'habitude d'utiliser le discours interne

Anxiété précompétitive

Méthodologie : Construction de différentes variables
- Anxiété générale : moyenne de tous les items
- Anxiété somatique : moyenne des items 1, 4, 7, 9, 10
- Anxiété cognitive : moyenne des items 2, 3, 5, 6, 8

Passion

Méthodologie :
- Passion harmonieuse : moyenne des items 1, 3, 5, 7, 9, 11
- Passion obsessive : moyenne des items 2, 4, 6, 8, 10, 12

Addiction

Méthodologie : Faire une addition du score de tous les items

Interprétation : Si le score est égal ou supérieur à 24 l'individu est « à risque de dépendance à l'exercice physique. Donc possibilité de créer 2 catégories ensuite « à risque » et non « à risque ».

Gestion des types de variables

Toutes les données récoltées du fichier Excel sont sous le type “Character”. Or pour nos représentations nous avons besoin d’avoir un typage de qualité pour pouvoir représenter nos variables. Ainsi pour certaines variables nous les avons mises sous la forme de factor et pour d’autres de numériques.

Pour les questions à choix multiples, nous sommes en train de créer une fonction globale nous permettant de mettre en colonne le nom de la réponse et de coder par 0 ou par 1 suivant si le coureur a coché cette case. Cette fonction sera ensuite rajouter dans notre fichier mise en forme et donc nous obtiendrons ses modifications dans notre jeu de données finales. Cette modification a aussi beaucoup d’intérêt pour réaliser des représentations graphiques.

Remarque

Nous avons remarqué lorsque nous traitons nos doublons que si l'on regroupe non pas que par leur numéro de dossier mais également par leur âge et sexe nous avions moins de doublons. En effet, 15 coureurs ont rempli deux fois le questionnaire avec un intervalle de temps ou leur anniversaire a du tombé. En effet, nous obtenons leur âge inscrit au précédent questionnaire + 1 an. Voici une visualisation pour mieux comprendre cette remarque.

Table 10: Visualisation des doublons dans notre jeu de données

Progress	Q64	Q1	Q2	Q3.1	Q3.2
100	BTK83111539	Homme	47	74	75
31	BTK83111539	Homme	48	74	75
100	DHP89151690	Homme	32	75	76
87	DHP89151690	Homme	33	75	76
100	IMJ55075863	Homme	47	71	74
17	IMJ55075863	Homme	57	71	74
100	JXG10263729	Homme	43	78	90
100	JXG10263729	Homme	44	80	85
100	LHK28906512	Homme	32	85	88
85	LHK28906512	Homme	31	85	89
100	REE83272089	Homme	65	72.2	75
31	REE83272089	Femme	65	72.2	75
100	UBO49849223	Homme	51	68	68
92	UBO49849223	Homme	52	68	68
100	UUU26690739	Homme	42	80	86
100	UUU26690739	Homme	43	80	83

Une seconde chose à remarquer dans ce tableau est l'individu avec le numéro de dossier IMJ55075863. En effet, ce coureur n'a pas un an de différence entre ses deux questionnaires mais 10 ans. L'hypothèse la plus logique ici est une faute de frappe réalisée par le coureur. Dans le jeu de données actuel, avec le code réalisé précédemment c'est la ligne où l'individu à 47 ans qui est gardée.