الجـــمهوريـــة الجزائريـــة الديمـــقراطيـــــة الشـعبيـة
ⵜⴰⵣⴰⵡⴰ ⵜⴰⵣⵣⴰⵢⵔⵜ ⵜⴰⵎⴰⵡⴰⵙⵜ ⵜⴰⵖⵡⵔⵉⵡⵜ
République Algérienne Démocratique et Populaire
وزارة التعليـــم العـــالي والبـــحث العلمـــي
ⴰⵢⵏⵉⵣ ⵏ ⵜⵓⵙⵎⵔⵙⵉⵏ ⵜⵏⵉⵣ ⴷ ⵜⵓⵔⵉ ⵜⵓⵙⵙⵏⵜ
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

ECOLE NATIONALE SUPÉRIEURE D'INFORMATIQUE

المدرسة الوطنية العليا للإعلام الآلي
ⴰⵢⵜⵓⵔⴰⵔ ⴰⵢⵏⵍⵍⵓ ⵜⵏⵉⵣ ⵏ ⵜⵓⵙⵏⵓⵎⵜⵉⵏ
École nationale Supérieure d'Informatique

Rapport TP ACP - ANAD

Option : Systèmes d'Information et Technologies (SIT2)

Sujet : ACP pour le PV de 1CS S1 de la promotion 2021

Auteurs :

- *ABDELKEBIR ACHRAF*
- *MAKHLOUFI AYMEN*

Proposé par :

- Mme HAMDAD Leila

Table des matières:

Table des matières: **2**
1. Introduction **3**
1.1. Objectif de l'analyse 3
1.2. Présentation des données et contexte de l'étude 3
2. Préparation des Données **4**
2.1. Fusion des fichiers de données 4
2.2. Prétraitement : Filtrage et nettoyage 5
2.3. Pondération des variables et Normalisation des données 7
- SYS1 : 5 7
- RES1 : 4 7
- ANUM : 4 7
- RO : 3 7
- ORG : 3 7
- LANG1 : 2 7
- IGL : 5 7
- THP : 4 7

3. Analyse en Composantes Principales **8**
Résultats globaux 8
Interprétation des variables 8
Projection des individus 8
Visualisation 9
Visualisation personnalisée 11
4. Gestion des données aberrantes **12**
4.1. Détection des valeurs aberrantes via Z-scores 12
4.1. Réalisation d'une nouvelle ACP 12
5. Analyse des résultats **14**
5.1. Projection des individus et des variables 14
5.2. Mise en évidence ma projection dans biplot 14
5.3. Visualisation par spécialité 16

1. Introduction

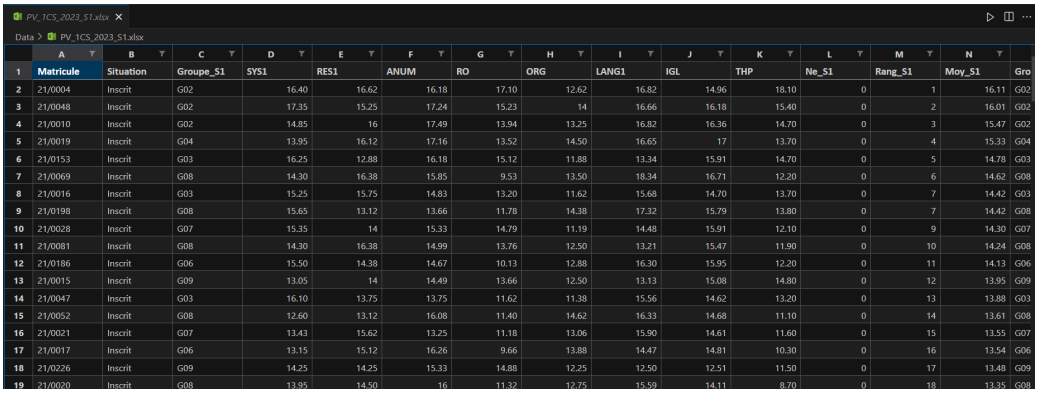
1.1. Objectif de l'analyse

L'objectif de cette analyse est d'appliquer une Analyse en Composantes Principales (ACP) sur les Procès-Verbaux (PVs) de délibération afin d'explorer les données, identifier les corrélations entre les modules et les étudiants, et vérifier la conformité des résultats aux spécialités attribuées. Cette étude permet également de détecter les données aberrantes, de les traiter et de visualiser les groupes formés par spécialité.

1.2. Présentation des données et contexte de l'étude

Les données utilisées dans cette étude proviennent de deux fichiers distincts :

1. **PV_1CS_2023_S1.xlsx** : Ce fichier contient les informations académiques des étudiants inscrits en première année de cycle supérieur (1CS), telles que les matricules, les groupes, les notes obtenues dans différents modules (ex. SYS1, RES1, ANUM), et d'autres détails comme le rang et la moyenne semestrielle.

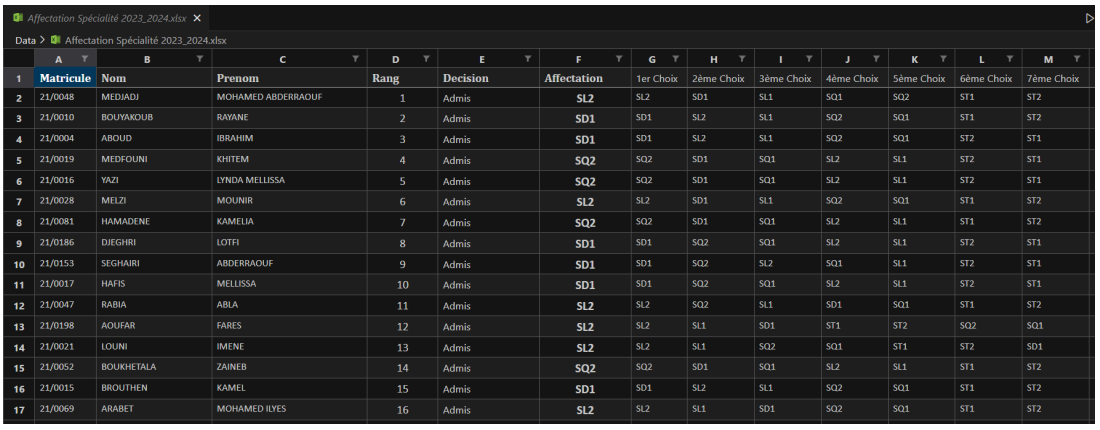


PV_1CS_2023_S1.xlsx

Data > PV_1CS_2023_S1.xlsx

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	
1	Matricule	Situation	Groupe_S1	SYS1	RES1	ANUM	RO	ORG	LANG1	IGL	THP	Ne_S1	Rang_S1	Moy_S1	Gro
2	21/0004	Inscrit	G02	16.40	16.62	16.18	17.10	12.62	16.82	14.96	18.10	0	1	16.11	G02
3	21/0048	Inscrit	G02	17.35	15.25	17.24	15.23	14	16.66	16.18	15.40	0	2	16.01	G02
4	21/0010	Inscrit	G02	14.85	16	17.49	13.94	13.25	16.82	16.36	14.70	0	3	15.47	G02
5	21/0019	Inscrit	G04	13.95	16.12	17.16	13.52	14.50	16.65	17	13.70	0	4	15.33	G04
6	21/0153	Inscrit	G03	16.25	12.88	16.18	15.12	11.88	13.34	15.91	14.70	0	5	14.78	G03
7	21/0069	Inscrit	G08	14.30	16.38	15.85	9.53	13.50	18.34	16.71	12.20	0	6	14.62	G08
8	21/0016	Inscrit	G03	15.25	15.75	14.83	13.20	11.62	15.68	14.70	13.70	0	7	14.42	G03
9	21/0198	Inscrit	G08	15.65	13.12	13.66	11.78	14.38	17.32	15.79	13.80	0	7	14.42	G08
10	21/0028	Inscrit	G07	15.35	14	15.33	14.79	11.19	14.48	15.91	12.10	0	9	14.30	G07
11	21/0081	Inscrit	G08	14.30	16.38	14.99	13.76	12.50	13.21	15.47	11.90	0	10	14.24	G08
12	21/0186	Inscrit	G06	15.50	14.38	14.67	10.13	12.88	16.30	15.95	12.20	0	11	14.13	G06
13	21/0015	Inscrit	G09	13.05	14	14.49	13.66	12.50	13.13	15.08	14.80	0	12	13.95	G09
14	21/0047	Inscrit	G03	16.10	13.75	13.75	11.62	11.38	15.56	14.62	13.20	0	13	13.88	G03
15	21/0052	Inscrit	G08	12.60	13.12	16.08	11.40	14.62	16.33	14.68	11.10	0	14	13.61	G08
16	21/0021	Inscrit	G07	13.43	15.62	13.25	11.18	13.06	15.90	14.61	11.60	0	15	13.55	G07
17	21/0017	Inscrit	G06	13.15	15.12	16.26	9.66	13.88	14.47	14.81	10.30	0	16	13.54	G06
18	21/0226	Inscrit	G09	14.25	14.25	15.33	14.88	12.25	12.50	12.51	11.50	0	17	13.48	G09
19	21/0020	Inscrit	G08	13.95	14.50	16	11.32	12.75	15.59	14.11	8.70	0	18	13.35	G08

2. **Affectation Spécialité 2023_2024.xlsx** : Ce fichier recense les spécialités attribuées aux étudiants en fonction de leurs choix et de leur rang. Il comprend également les décisions d'admission et les préférences exprimées par les étudiants pour diverses spécialités.



Affectation Spécialité 2023_2024.xlsx

Data > Affectation Spécialité 2023_2024.xlsx

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Matricule	Nom	Prenom	Rang	Decision	Affectation	1er Choix	2ème Choix	3ème Choix	4ème Choix	5ème Choix	6ème Choix	7ème Choix
2	21/0048	MEDJADJ	MOHAMED ABDERRAOUF	1	Admis	SL2	SL2	SD1	SL1	SQ1	SQ2	ST1	ST2
3	21/0010	BOUYAKOUB	RAYANE	2	Admis	SD1	SD1	SL2	SL1	SQ2	SQ1	ST1	ST2
4	21/0004	ABOUD	IBRAHIM	3	Admis	SD1	SD1	SL2	SL1	SQ2	SQ1	ST2	ST1
5	21/0019	MEDFOUNI	KHITEM	4	Admis	SQ2	SQ2	SD1	SQ1	SL2	SL1	ST2	ST1
6	21/0016	YAZI	LYNDA MELLISSA	5	Admis	SQ2	SQ2	SD1	SQ1	SL2	SL1	ST2	ST1
7	21/0028	MELZI	MOUNIR	6	Admis	SL2	SL2	SD1	SL1	SQ2	SQ1	ST1	ST2
8	21/0081	HAMADENE	KAMELIA	7	Admis	SQ2	SQ2	SD1	SQ1	SL2	SL1	ST1	ST2
9	21/0186	DJEGHRI	LOTFI	8	Admis	SD1	SD1	SQ2	SQ1	SL2	SL1	ST2	ST1
10	21/0153	SEGHAIRI	ABDERRAOUF	9	Admis	SD1	SD1	SQ2	SL2	SQ1	SL1	ST2	ST1
11	21/0017	HAFIS	MELLISSA	10	Admis	SD1	SD1	SQ2	SQ1	SL2	SL1	ST2	ST1
12	21/0047	RABIA	ABLA	11	Admis	SL2	SL2	SQ2	SL1	SD1	SQ1	ST1	ST2
13	21/0198	AOUFAR	FARES	12	Admis	SL2	SL2	SL1	SD1	ST1	ST2	SQ2	SQ1
14	21/0021	LOUNI	IMENE	13	Admis	SL2	SL2	SL1	SQ2	SQ1	ST1	ST2	SD1
15	21/0052	BOUKHETALA	ZAINEB	14	Admis	SQ2	SQ2	SD1	SQ1	SL2	SL1	ST1	ST2
16	21/0015	BROUTHEN	KAMEL	15	Admis	SD1	SD1	SL2	SL1	SQ2	SQ1	ST1	ST2
17	21/0069	ARABET	MOHAMED ILYES	16	Admis	SL2	SL2	SL1	SD1	SQ2	SQ1	ST1	ST2

Ces données permettent d'explorer la relation entre les performances académiques des étudiants et leur spécialité attribuée, de visualiser les groupes formés, et d'examiner si les résultats observés sont cohérents avec les compétences requises pour chaque spécialité.

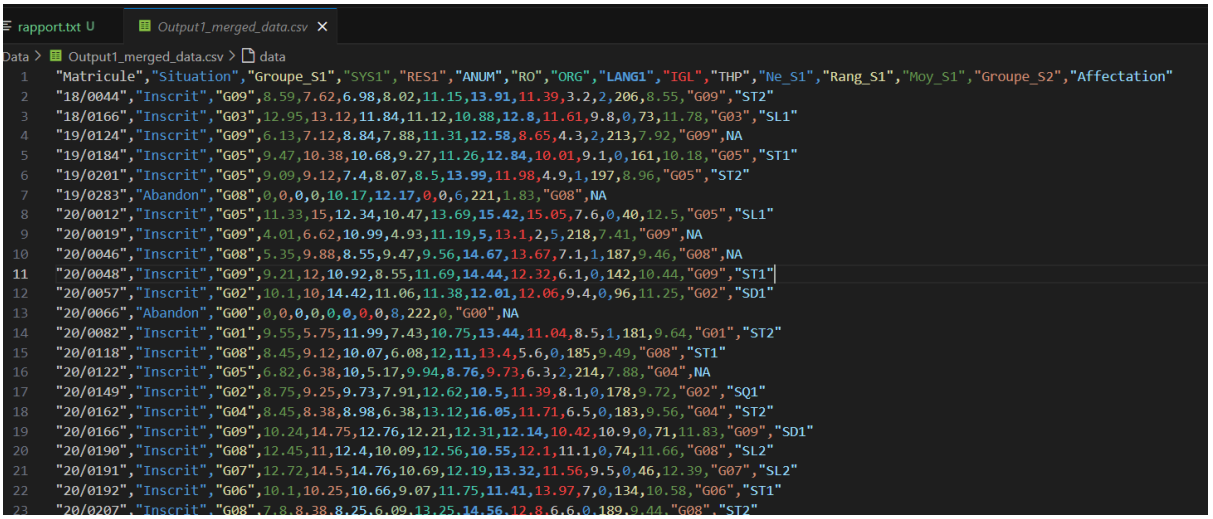
2. Préparation des Données

2.1. Fusion des fichiers de données

Pour effectuer l'analyse, les deux fichiers de données ont été fusionnés en utilisant le matricule comme clé commune. Cette opération a permis d'associer les informations académiques des étudiants (notes, groupes, etc.) avec leur spécialité attribuée et leurs choix.

```r
# Charger les bibliothèques nécessaires
library(FactoMineR)
library(factoextra)
library(ggplot2)
library(openxlsx)

# Étape 1 : Charger les données ----------------------------------------------------------------------
# Récupérer le répertoire de travail actuel
current_dir <- getwd()

# Construire automatiquement le chemin du fichier en utilisant getwd()
file_path1 <- paste0(current_dir, "/Data/Affectation Spécialité 2023_2024.xlsx")
file_path2 <- paste0(current_dir, "/Data/PV_1CS_2023_S1.xlsx")

# Charger les données depuis les fichiers
data <- read.xlsx(file_path2)
affectation_data <- read.xlsx(file_path1)

head(data)
head(affectation_data)


# Fusionner les deux jeux de données sur "Matricule" avec left join
merged_data <- merge(data, affectation_data[, c("Matricule", "Affectation")],
    by = "Matricule", all.x = TRUE
)

# Afficher un aperçu des données fusionnées
head(merged_data)

# Save marged_data as csv in data folder
write.csv(merged_data, file = "Data/Output1_merged_data.csv", row.names = FALSE)

```

La fusion a été réalisée en tant que jointure gauche (left join) afin de conserver tous les étudiants présents dans le fichier académique, même si leur spécialité n'était pas initialement renseignée. Cela garantit que l'analyse inclut tous les étudiants inscrits. le résultat de cette opération a été enregistré dans le fichier ***Output1_merged_data.csv.***



```
rapport.txt U   Output1_merged_data.csv
Data > Output1_merged_data.csv > data
"Matricule","Situation","Groupe_S1","SYS1","RES1","ANUM","RO","ORG","LANG1","IGL","THP","Ne_S1","Rang_S1","Moy_S1","Groupe_S2","Affectation"
"18/0044","Inscrit","G09",8.59,7.62,6.98,8.02,11.15,13.91,11.39,3.2,2,206,8.55,"G09","ST2"
"18/0166","Inscrit","G03",12.95,13.12,11.84,11.12,10.88,12.8,11.61,9.8,0,73,11.78,"G03","SL1"
"19/0124","Inscrit","G09",6.13,7.12,8.84,7.88,11.31,12.58,8.65,4.3,2,213,7.92,"G09",NA
"19/0184","Inscrit","G05",9.47,10.38,10.68,9.27,11.26,12.84,10.01,9.1,0,161,10.18,"G05","ST1"
"19/0201","Inscrit","G05",9.09,9.12,7.4,8.07,8.5,13.99,11.98,4.9,1,197,8.96,"G05","ST2"
"19/0283","Abandon","G08",0,0,0,0,10.17,12.17,0,0,6,221,1.83,"G08",NA
"20/0012","Inscrit","G05",11.33,15,12.34,10.47,13.69,15.42,15.05,7.6,0,40,12.5,"G05","SL1"
"20/0019","Inscrit","G09",4.01,6.62,10.99,4.93,11.19,5,13.1,2,5,218,7.41,"G09",NA
"20/0046","Inscrit","G08",5.35,9.88,8.55,9.47,9.56,14.67,13.67,7.1,1,187,9.46,"G08",NA
"20/0048","Inscrit","G09",9.21,12,10.92,8.55,11.69,14.44,12.32,6.1,0,142,10.44,"G09","ST1"
"20/0057","Inscrit","G02",10.1,10,14.42,11.06,11.38,12.01,12.06,9.4,0,96,11.25,"G02","SD1"
"20/0066","Abandon","G00",0,0,0,0,0,0,0,8,222,0,"G00",NA
"20/0082","Inscrit","G01",9.55,5.75,11.99,7.43,10.75,13.44,11.04,8.5,1,181,9.64,"G01","ST2"
"20/0118","Inscrit","G08",8.45,9.12,10.07,6.08,12,11,13.4,5.6,0,185,9.49,"G08","ST1"
"20/0122","Inscrit","G05",6.82,6.38,10,5.17,9.94,8.76,9.73,6.3,2,214,7.88,"G04",NA
"20/0149","Inscrit","G02",8.75,9.25,9.73,7.91,12.62,10.5,11.39,8.1,0,178,9.72,"G02","SQ1"
"20/0162","Inscrit","G04",8.45,8.38,8.98,6.38,13.12,16.05,11.71,6.5,0,183,9.56,"G04","ST2"
"20/0166","Inscrit","G09",10.24,14.75,12.76,12.21,12.31,12.14,10.42,10.9,0,71,11.83,"G09","SD1"
"20/0190","Inscrit","G08",12.45,11,12.4,10.09,12.56,10.55,12.1,11.1,0,74,11.66,"G08","SL2"
"20/0191","Inscrit","G07",12.72,14.5,14.76,10.69,12.19,13.32,11.56,9.5,0,46,12.39,"G07","SL2"
"20/0192","Inscrit","G06",10.1,10.25,10.66,9.07,11.75,11.41,13.97,7,0,134,10.58,"G06","ST1"
"20/0207","Inscrit","G08",7.8,8.38,8.25,6.09,13.25,14.56,12.8,6.6,0,189,9.44,"G08","ST2"
```

2.2. Prétraitement : Filtrage et nettoyage

Le prétraitement des données a consisté à :

1. **Filtrer les données** :
    - Seuls les étudiants en situation *"Inscrit"* ont été conservés. Les cas d'abandon ou de congé académique ont été exclus.

```r
# Étape 3 : Prétraitement des données ----------------------------------------------------------

# Garder uniquement les lignes où la colonne "Situation" est égale à "Inscrit"
# Donc supprimer "Abandon" et "Congé académique (année blanche) pour raisons médicales"
filtered_data <- merged_data[merged_data$Situation == "Inscrit", ]
```

2. **Nettoyage** :
    - Les colonnes non pertinentes pour l'analyse, telles que *"Situation"*, *"Groupe_S1"*, *"Rang_S1"*, *"Moy_S1"*, et *"Groupe_S2"*, ont été supprimées.
    - Les valeurs manquantes dans la colonne *"Affectation"* ont été remplacées par "NonAdmis2CS".
    - Le résultat a été enregistré dans le fichier ***Output2_preprocessed_data.csv.***

```r
# Retirer les colonnes "Situation", "Groupe_S1", "Rang_S1", "Moy_S1", et "Groupe_S2" car elles ne sont pas pertinentes pour l'ACP
filtered_data <- filtered_data[, !colnames(filtered_data) %in% c("Situation", "Groupe_S1", "Rang_S1", "Moy_S1", "Groupe_S2")]

# Remplir les valeurs manquantes dans la colonne "Affectation" avec "NonAdmis2CS"
filtered_data$Affectation[is.na(filtered_data$Affectation)] <- "NonAdmis2CS"

# Afficher un aperçu des données traitées pour s'assurer que les étapes ont bien été exécutées
head(filtered_data)

# Si vous souhaitez enregistrer les données dans un fichier CSV pour des usages ultérieurs
write.csv(filtered_data, file = "Data/Output2_preprocessed_data.csv", row.names = FALSE)
```

3. **Création d'index unique** :
    - Un identifiant unique a été généré en combinant le matricule et la spécialité pour faciliter le suivi des individus dans l'analyse.
    - Le résultat a été enregistré dans le fichier ***Output2_preprocessed_data.csv.***

```r

# Création d'un index unique pour chaque ligne du dataframe ----------------------------------------------------
# Combiner les colonnes "Matricule" et "Affectation" pour créer une nouvelle colonne "Index"
# Utilisation de paste pour concaténer les deux colonnes avec un underscore comme séparateur
filtered_data$Index <- paste(filtered_data$Matricule, filtered_data$Affectation, sep = "_")


# La colonne "Index" devient l'identifiant unique pour chaque ligne du dataframe pour nous aider aprés dans la vis
rownames(filtered_data) <- filtered_data$Index


# Supprimer les colonnes "Matricule", "Affectation", "Index" et "Ne_S1" si elles ne sont plus nécessaires
filtered_data <- filtered_data[, !colnames(filtered_data) %in% c("Matricule", "Affectation", "Index", "Ne_S1")]

# Afficher un aperçu des données pour s'assurer que l'index est bien défini
head(filtered_data)

# Enregistrer le dataframe modifié dans un fichier CSV avec row.names=TRUE pour inclure l'index
write.csv(filtered_data, file = "Data/Output3_indexed_data.csv", row.names = TRUE)

```

Data > Output3_indexed_data.csv > data

1	""	"SYS1"	"RES1"	"ANUM"	"RO"	"ORG"	"LANG1"	"IGL"	"THP"
2	"18/0044_ST2"	8.59	7.62	6.98	8.02	11.15	13.91	11.39	3.2
3	"18/0166_SL1"	12.95	13.12	11.84	11.12	10.88	12.8	11.61	9.8
4	"19/0124_NonAdmis2CS"	6.13	7.12	8.84	7.88	11.31	12.58	8.65	4.3
5	"19/0184_ST1"	9.47	10.38	10.68	9.27	11.26	12.84	10.01	9.1
6	"19/0201_ST2"	9.09	9.12	7.4	8.07	8.5	13.99	11.98	4.9
7	"20/0012_SL1"	11.33	15	12.34	10.47	13.69	15.42	15.05	7.6
8	"20/0019_NonAdmis2CS"	4.01	6.62	10.99	4.93	11.19	5	13.1	2
9	"20/0046_NonAdmis2CS"	5.35	9.88	8.55	9.47	9.56	14.67	13.67	7.1
10	"20/0048_ST1"	9.21	12	10.92	8.55	11.69	14.44	12.32	6.1
11	"20/0057_SD1"	10.1	10	14.42	11.06	11.38	12.01	12.06	9.4
12	"20/0082_ST2"	9.55	5.75	11.99	7.43	10.75	13.44	11.04	8.5
13	"20/0118_ST1"	8.45	9.12	10.07	6.08	12	11	13.4	5.6
14	"20/0122_NonAdmis2CS"	6.82	6.38	10	5.17	9.94	8.76	9.73	6.3
15	"20/0149_SQ1"	8.75	9.25	9.73	7.91	12.62	10.5	11.39	8.1
16	"20/0162_ST2"	8.45	8.38	8.98	6.38	13.12	16.05	11.71	6.5

Ces étapes ont permis d'obtenir un jeu de données propre et cohérent, prêt pour les analyses statistiques.

2.3. Pondération des variables et Normalisation des données

La pondération des variables a été effectuée afin d'ajuster l'importance relative de chaque module dans l'analyse. Les coefficients attribués à chaque module sont les suivants:

- SYS1 : 5
- RES1 : 4
- ANUM : 4
- RO : 3
- ORG : 3
- LANG1 : 2
- IGL : 5
- THP : 4

Ces coefficients ont été utilisés pour multiplier les valeurs des colonnes correspondantes, modifiant ainsi l'impact de chaque module dans l'analyse finale.

```r

# Définir les coefficients des modules
coefficients <- c(SYS1 = 5, RES1 = 4, ANUM = 4, RO = 3, ORG = 3, LANG1 = 2, IGL = 5, THP = 4)

# Appliquer la pondération aux colonnes correspondantes
for (module in names(coefficients)) {
  filtered_data[[module]] <- filtered_data[[module]] * coefficients[module]
}

```

Ensuite, les données ont été normalisées à l'aide de la fonction `scale()`, ce qui permet de standardiser les variables, les transformant en une échelle commune de manière à ce qu'elles aient toutes une moyenne de 0 et un écart-type de 1. Cette étape est cruciale pour l'ACP, car elle évite que les variables avec des échelles différentes dominent l'analyse.

```r

# Normaliser les données pondérées
normalized_data <- scale(filtered_data)

head(normalized_data)

# Enregistrer les données normalisées dans un fichier CSV
write.csv(normalized_data, file = "Data/Output4_normalized_data.csv", row.names = TRUE)

```

Data > Output4_normalized_data.csv > data

```
"","SYS1","RES1","ANUM","RO","ORG","LANG1","IGL","THP"
"18/0044_ST2",-0.938349316514097,-1.38974235976238,-2.09145294556657,-0.575687368299996,-0.309455587417951,0.464680891970651,-0.940313384918
"18/0166_SL1",0.943764907902117,0.803938833245966,0.0141793420038414,0.734526825054464,-0.536473765664,0.0126790046205894,-0.804113570972246,0.46
"19/0124_NonAdmis2CS",-2.00027614955627,-1.58916792276313,-1.28559367501493,-0.634858331870843,-0.174926296605159,-0.076906955034378,-2.63662015860
"19/0184_ST1",-0.558473051035595,-0.288913251998189,-0.488399557910085,-0.0473751935602948,-0.216966699984156,0.0289673609214922,-1.79465767239737,
"19/0201_ST2",-0.722510529310403,-0.791465670760101,-1.90948472318394,-0.554554881310408,-2.53759696650484,0.497257604572457,-0.575050247517684,-1.
"20/0012_SL1",0.244447237362147,1.55377895012882,0.230808178173637,0.459804494189819,1.82619690423514,1.07956634232974,1.32555624709178,-0.40157238
"20/0019_NonAdmis2CS",-2.91543260729994,-1.78859348576389,-0.354089679484811,-1.88167506425654,-0.275823264714753,-3.16355047405552,0.1183306234799
"20/0046_NonAdmis2CS",-2.33698465759404,-0.488338814998948,-1.41123839999341,0.0371547543980581,-1.646340414870008,0.77415966168781,0.471211959612600
"20/0048_ST1",-0.670709220381516,0.357225572124268,-0.384417716548583,-0.351683006210362,0.144580769075225,0.680501612957616,-0.364559625964845,-0.
"20/0057_SD1",-0.286516179158941,-0.440476679878766,1.13198413663999,0.709167840666958,-0.116069731874561,-0.309016032322247,-0.525523042446428,0.3
"20/0082_ST2",-0.523938845083004,-2.13559396538521,0.0791679928547802,-0.825050714777135,-0.645778814449935,0.273292705435039,-1.15699490710495,-0.
"20/0118_ST1",-0.998784176931132,-0.791465670760101,-0.752686738037235,-1.39562786349601,0.405231270025012,-0.720297028920051,0.304057642497116,-1.
"20/0122_NonAdmis2CS",-1.70241862321518,-1.88431775600426,-0.783014775101007,-1.78023912670652,-1.32683333491897,-1.63244498177063,-1.96800289014677
"20/0149_SQ1",-0.869280904608915,-0.739615024379904,-0.899994346632696,-0.622178839670 9,0.926532271924586,-0.923901482681 34,-0.940313384918 2,-0.20
"20/0162_ST2",-0.998784176931132,-1.08661550400122,-1.22493760088739,-1.26883294155848,1.34693630571456,1.33610795406897,-0.742204564633175,-0.8336
"20/0166_SD1",-0.226081318741906,1.45406616862844,0.412776400556265,1.19521504142748,0.665881770974799,-0.256078874344312,-1.54083074640718,0.89474
"20/0190_SL2",0.727926120698423,-0.0416255538772493,0.256803638514013,0.299197593068949,0.876083787869788,-0.903541037305211,-0.5007594399108,0.973
"20/0191_SL2",0.844479065788418,1.35435338712806,1.27929174523545,0.552787436944006,0.564984802865204,0.22442763653233,-0.83506807414178,0.34479158
"20/0192_ST1",-0.286516179158941,-0.340763898378387,-0.497064711356876,-0.131905141518647,0.195029253130023,-0.553341376835794,0.656938978629819,-0
```

3. Analyse en Composantes Principales

L'Analyse en Composantes Principales (ACP) a permis d'identifier les dimensions les plus significatives dans les données. Voici un résumé des résultats :

Résultats globaux

```
> # Appliquer l'ACP sur les données normalisées
> acp_result <- PCA(normalized_data, graph = FALSE)
> summary(acp_result)

Call:
PCA(X = normalized_data, graph = FALSE)


Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7
Variance               4.492   0.944   0.714   0.499   0.379   0.361   0.323
% of var.             56.153  11.799   8.921   6.241   4.737   4.514   4.036
Cumulative % of var.  56.153  67.952  76.874  83.115  87.852  92.366  96.402
                       Dim.8
Variance               0.288
% of var.              3.598
Cumulative % of var. 100.000

Individuals (the 10 first)
                      Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2
18/0044_ST2          | 3.648 | -3.050  0.941  0.699 |  1.353  0.881  0.137 |
18/0166_SL1          | 1.800 |  0.735  0.055  0.167 | -1.060  0.541  0.347 |
19/0124_NonAdmis2CS  | 4.306 | -3.832  1.486  0.792 |  0.525  0.133  0.015 |
19/0184_ST1          | 1.989 | -1.173  0.139  0.348 | -0.440  0.093  0.049 |
19/0201_ST2          | 3.785 | -2.808  0.798  0.550 | -0.163  0.013  0.002 |
20/0012_SL1          | 3.033 |  2.029  0.417  0.448 |  1.779  1.525  0.344 |
20/0019_NonAdmis2CS  | 5.690 | -4.523  2.070  0.632 | -0.095  0.004  0.000 |
20/0046_NonAdmis2CS  | 3.411 | -1.934  0.379  0.322 |  0.357  0.061  0.011 |
20/0048_ST1          | 1.568 | -0.710  0.051  0.205 |  0.909  0.398  0.336 |
20/0057_SD1          | 1.597 |  0.221  0.005  0.019 | -0.809  0.315  0.256 |
                      Dim.3    ctr   cos2
18/0044_ST2          -0.555  0.196  0.023 |
18/0166_SL1          -0.412  0.108  0.052 |
19/0124_NonAdmis2CS -0.041  0.001  0.000 |
19/0184_ST1          -0.147  0.014  0.005 |
19/0201_ST2          -2.187  3.047  0.334 |
20/0012_SL1           0.558  0.198  0.034 |
20/0019_NonAdmis2CS  2.014  2.582  0.125 |
20/0046_NonAdmis2CS -1.710  1.861  0.251 |
20/0048_ST1          -0.356  0.081  0.051 |
20/0057_SD1           0.121  0.009  0.006 |
```

```
Variables
                    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3
SYS1                | 0.858 16.372  0.735 | -0.090  0.860  0.008 | -0.015
RES1                | 0.842 15.796  0.710 | -0.123  1.594  0.015 |  0.007
ANUM                | 0.800 14.235  0.639 | -0.171  3.103  0.029 | -0.001
RO                  | 0.793 14.000  0.629 | -0.278  8.189  0.077 | -0.023
ORG                 | 0.526  6.147  0.276 |  0.556 32.755  0.309 |  0.613
LANG1               | 0.522  6.065  0.272 |  0.586 36.398  0.344 | -0.581
IGL                 | 0.770 13.198  0.593 |  0.249  6.560  0.062 | -0.008
RES1                | 0.842 15.796  0.710 | -0.123  1.594  0.015 |  0.007
ANUM                | 0.800 14.235  0.639 | -0.171  3.103  0.029 | -0.001
RO                  | 0.793 14.000  0.629 | -0.278  8.189  0.077 | -0.023
ORG                 | 0.526  6.147  0.276 |  0.556 32.755  0.309 |  0.613
LANG1               | 0.522  6.065  0.272 |  0.586 36.398  0.344 | -0.581
RES1                | 0.842 15.796  0.710 | -0.123  1.594  0.015 |  0.007
ANUM                | 0.800 14.235  0.639 | -0.171  3.103  0.029 | -0.001
RO                  | 0.793 14.000  0.629 | -0.278  8.189  0.077 | -0.023
ORG                 | 0.526  6.147  0.276 |  0.556 32.755  0.309 |  0.613
RES1                | 0.842 15.796  0.710 | -0.123  1.594  0.015 |  0.007
ANUM                | 0.800 14.235  0.639 | -0.171  3.103  0.029 | -0.001
RO                  | 0.793 14.000  0.629 | -0.278  8.189  0.077 | -0.023
ORG                 | 0.526  6.147  0.276 |  0.556 32.755  0.309 |  0.613
RES1                | 0.842 15.796  0.710 | -0.123  1.594  0.015 |  0.007
ANUM                | 0.800 14.235  0.639 | -0.171  3.103  0.029 | -0.001
RES1                | 0.842 15.796  0.710 | -0.123  1.594  0.015 |  0.007
ANUM                | 0.800 14.235  0.639 | -0.171  3.103  0.029 | -0.001
RES1                | 0.842 15.796  0.710 | -0.123  1.594  0.015 |  0.007
RES1                | 0.842 15.796  0.710 | -0.123  1.594  0.015 |  0.007
ANUM                | 0.800 14.235  0.639 | -0.171  3.103  0.029 | -0.001
RO                  | 0.793 14.000  0.629 | -0.278  8.189  0.077 | -0.023
ORG                 | 0.526  6.147  0.276 |  0.556 32.755  0.309 |  0.613
LANG1               | 0.522  6.065  0.272 |  0.586 36.398  0.344 | -0.581
IGL                 | 0.770 13.198  0.593 |  0.249  6.560  0.062 | -0.008
THP                 | 0.798 14.187  0.637 | -0.315 10.542  0.100 |  0.017
                     ctr   cos2
SYS1                0.030  0.000 |
RES1                0.006  0.000 |
ANUM                0.000  0.000 |
RO                  0.074  0.001 |
ORG                52.602  0.375 |
LANG1              47.238  0.337 |
IGL                 0.010  0.000 |
```

- **Variance expliquée :**
  - La première composante principale (Dim.1) explique 56,15 % de la variance totale.
  - La deuxième composante (Dim.2) explique 11,80 %.
  - Ensemble, les deux premières composantes expliquent **67,95 %** de la variance, offrant une bonne réduction dimensionnelle.

Interprétation des variables

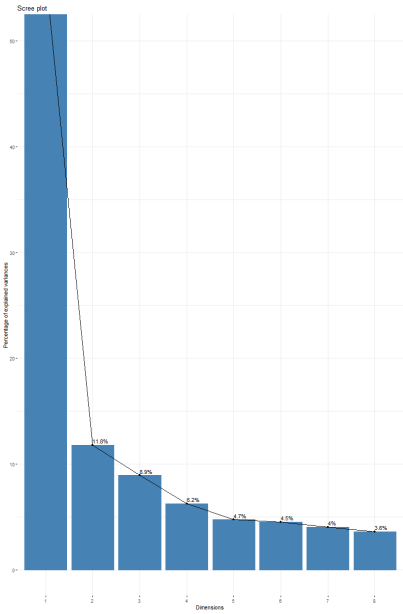
- Les variables **SYS1**, **RES1**, **ANUM**, **RO**, **THP**, et **IGL** contribuent fortement à **Dim.1**.
- Les variables **ORG** et **LANG1** sont majoritairement associées à **Dim.2**, avec une contribution significative et des cos2 élevés.

Projection des individus

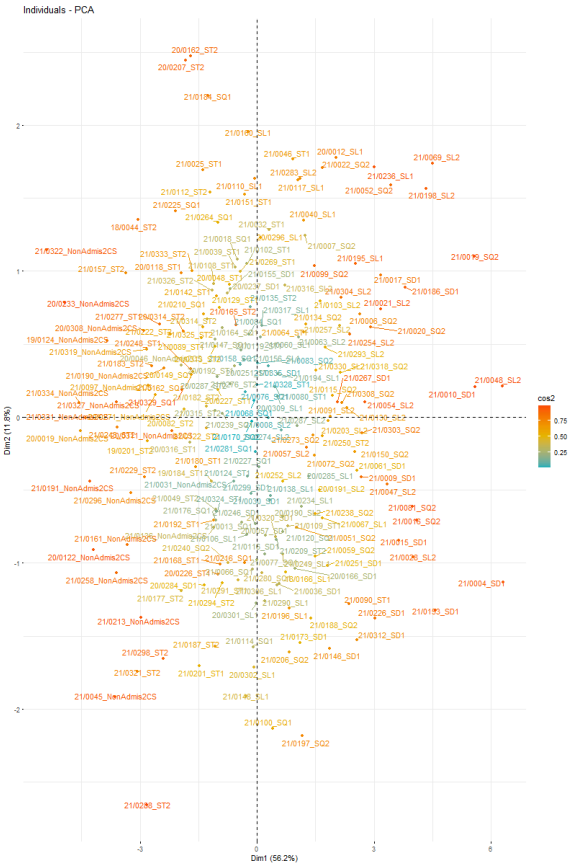
- Les individus ayant des contributions élevées sur **Dim.1** sont principalement différenciés par leurs performances dans les modules techniques (exemple : **SYS1**, **RES1**, etc.).
- Les individus projetés sur **Dim.2** reflètent des compétences linguistiques et organisationnelles.

Visualisation

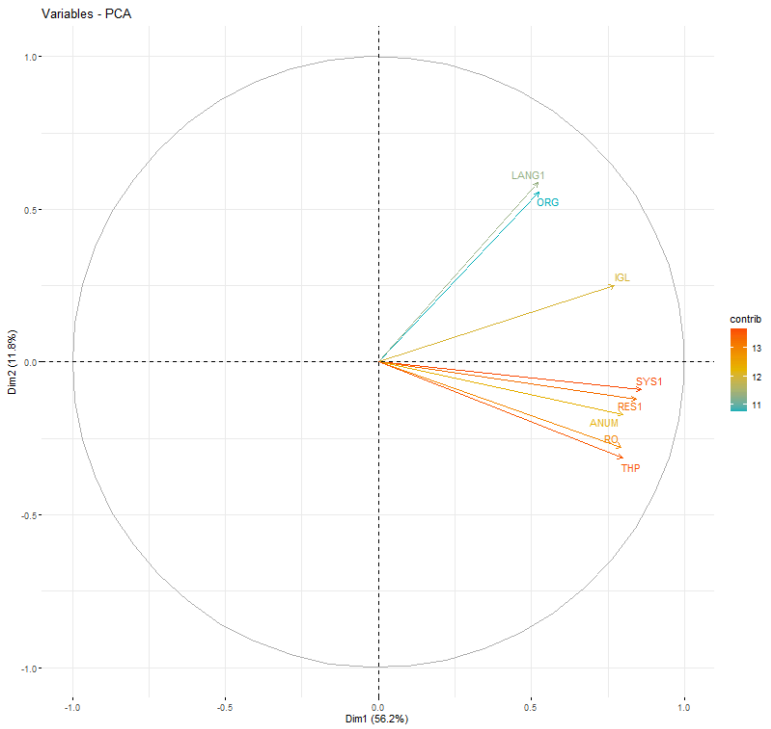
- **Diagramme des éboulis :** Montre une nette diminution après les deux premières composantes, justifiant leur utilisation.



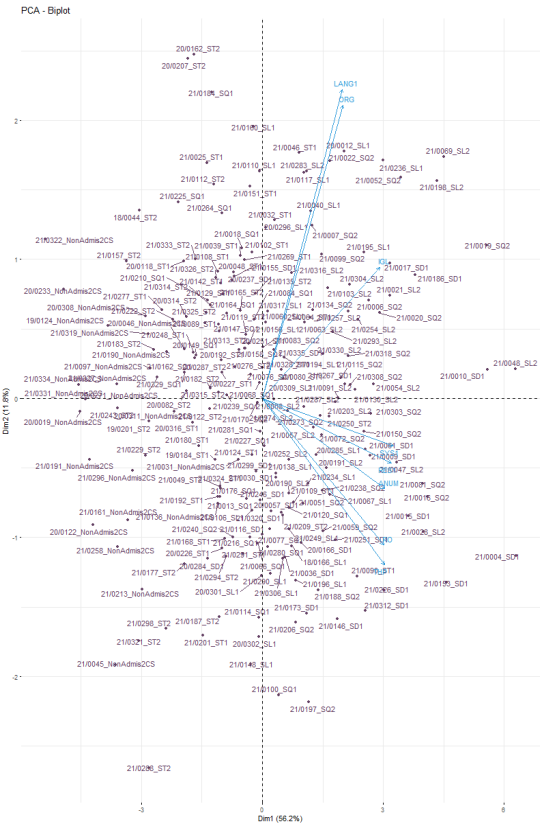
- **Nuage des individus :** Aide à identifier des groupes d'étudiants ayant des performances similaires.



- **Nuage des variables :** Visualise la relation entre les modules et leurs contributions aux axes principaux.



- **Biplot** : Une combinaison des deux graphiques précédents, le biplot superpose les individus et les variables sur le même plan.



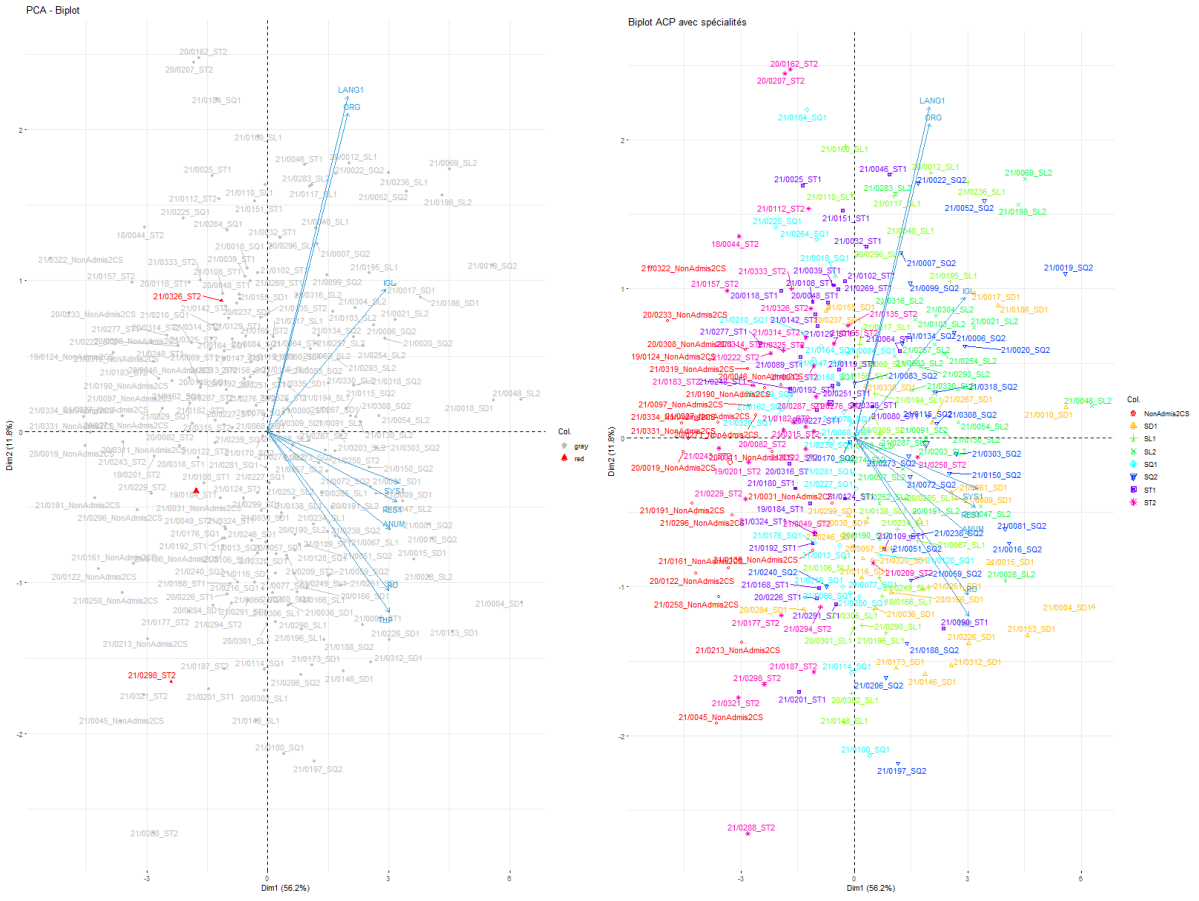
- **Biplot** : Une combinaison des deux graphiques précédents, le biplot superpose les individus et les variables sur le même plan.

Les résultats de l'ACP sont sauvegardés dans le fichier ***`Data/Result1_acp_result.rds`*** pour des analyses futures.

Visualisation personnalisée

Le **biplot** a été enrichi pour offrir une analyse plus détaillée :

- Les individus sont différenciés par des couleurs :
  - Mon binôme et moi-même sont mis en évidence en **rouge** pour une meilleure identification.
    - **ABDELKEBIR Achraf** id = **21/0298_ST2**
    - **MAKHLOUFI Aymen** id = **21/0326_ST2**
  - Les autres individus apparaissent en **gris** par défaut.
- Une seconde version colore les individus selon leur spécialité, offrant une vue globale des regroupements par filière.



Ces visualisations facilitent l'interprétation des contributions individuelles et des relations entre les variables tout en soulignant les distinctions spécifiques.

4. Gestion des données aberrantes

La gestion des données aberrantes est une étape cruciale dans l'analyse de données, notamment pour garantir la qualité des résultats obtenus. Dans cette section, nous avons appliqué des techniques pour identifier et éliminer les valeurs aberrantes, en particulier à travers l'utilisation des scores Z, suivie d'une nouvelle réalisation de l'Analyse en Composantes Principales (ACP) sur les données nettoyées.

4.1. Détection des valeurs aberrantes via Z-scores

Pour détecter les valeurs aberrantes, nous avons calculé les Z-scores des individus sur les deux premières composantes principales issues de l'ACP. Un Z-score élevé indique une valeur éloignée de la moyenne, ce qui peut signifier qu'il s'agit d'une donnée aberrante. Nous avons utilisé un seuil de 1,75 pour identifier les individus présentant des scores anormaux.

```
> # Étape 6 : Gestion des données aberrantes --------------------------------$
>
> # Identifier les individus avec des scores anormaux sur les composantes prin$
> outlier_scores <- acp_result$ind$coord[, 1:2] # Scores des deux premières co$
> z_scores_outliers <- scale(outlier_scores) # Calcul des z-scores pour les sc$
>
>
> is_outlier <- apply(z_scores_outliers, 1, function(x) any(abs(x) > 1.75))
>
>
> # Extraire les indices des individus aberrants
> outlier_indices <- which(is_outlier)
>
> # Afficher les individus identifiés comme aberrants
> outliers <- rownames(normalized_data)[outlier_indices]
> print("Individus aberrants détectés :")
[1] "Individus aberrants détectés :"
> print(outliers)
 [1] "19/0124_NonAdmis2CS" "20/0012_SL1"         "20/0019_NonAdmis2CS"
 [4] "20/0122_NonAdmis2CS" "20/0162_ST2"         "20/0207_ST2"
 [7] "20/0233_NonAdmis2CS" "20/0302_SL1"         "21/0004_SD1"
[10] "21/0010_SD1"         "21/0016_SQ2"         "21/0019_SQ2"
[13] "21/0022_SQ2"         "21/0028_SL2"         "21/0045_NonAdmis2CS"
[16] "21/0046_ST1"         "21/0048_SL2"         "21/0069_SL2"
[19] "21/0081_SQ2"         "21/0100_SQ1"         "21/0148_SL1"
[22] "21/0153_SD1"         "21/0160_SL1"         "21/0184_SQ1"
[25] "21/0186_SD1"         "21/0191_NonAdmis2CS" "21/0197_SQ2"
[28] "21/0198_SL2"         "21/0201_ST1"         "21/0236_SL1"
[31] "21/0288_ST2"         "21/0321_ST2"         "21/0322_NonAdmis2CS"
[34] "21/0331_NonAdmis2CS" "21/0334_NonAdmis2CS"
>
> # Supprimer les individus aberrants des données normalisées
> cleaned_data <- normalized_data[!rownames(normalized_data) %in% outliers, ]
>
```

4.1. Réalisation d'une nouvelle ACP

Après avoir éliminé les individus aberrants, nous avons réappliqué l'ACP sur les données nettoyées. Le résultat montre une distribution des variances similaire à celle obtenue précédemment, avec les premières composantes principales expliquant la majorité de la

variance. En effet, les deux premières composantes expliquent respectivement 45,68% et 11,90% de la variance totale, soit un total cumulé de 57,58%.

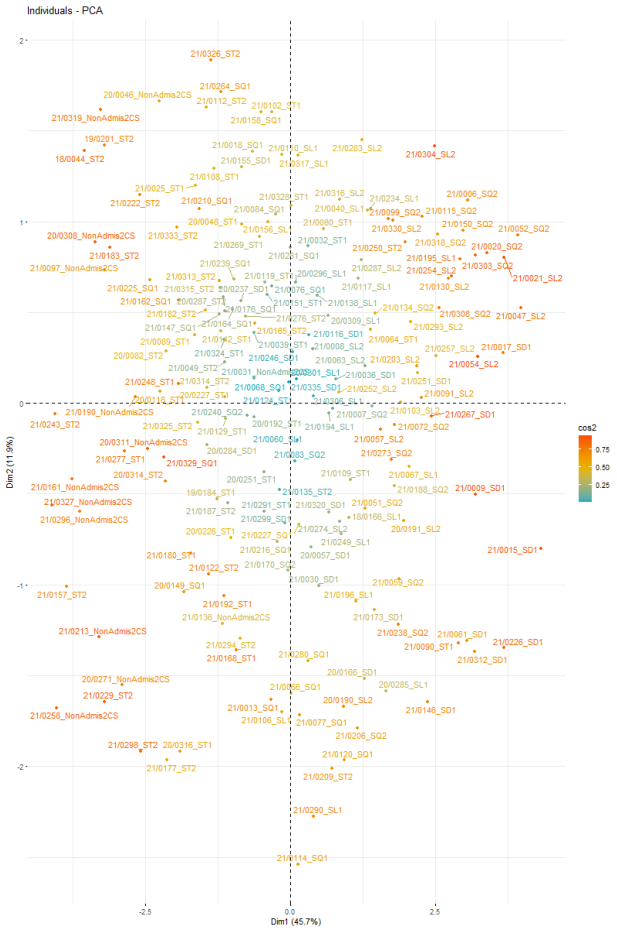
```
> # Réappliquer l'ACP sur les données nettoyées
> acp_result_cleaned <- PCA(cleaned_data, graph = FALSE)
> summary(acp_result_cleaned)

Call:
PCA(X = cleaned_data, graph = FALSE)


Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7
Variance               3.654   0.952   0.875   0.700   0.530   0.497   0.403
% of var.             45.678  11.900  10.943   8.755   6.620   6.211   5.038
Cumulative % of var.  45.678  57.579  68.522  77.277  83.897  90.107  95.146
                       Dim.8
Variance               0.388
% of var.              4.854
Cumulative % of var. 100.000

Individuals (the 10 first)
                        Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2
18/0044_ST2          |  4.157 | -3.561  1.875  0.734 |  1.389  1.095  0.112 |
18/0166_SL1          |  2.107 |  1.008  0.150  0.229 | -0.631  0.226  0.090 |
19/0184_ST1          |  2.335 | -1.265  0.237  0.294 | -0.529  0.159  0.051 |
19/0201_ST2          |  4.159 | -3.203  1.517  0.593 |  1.421  1.147  0.117 |
20/0046_NonAdmis2CS  |  3.880 | -2.269  0.761  0.342 |  1.663  1.570  0.184 |
20/0048_ST1          |  1.810 | -0.838  0.104  0.214 |  0.985  0.551  0.296 |
20/0057_SD1          |  1.846 |  0.357  0.019  0.037 | -0.793  0.357  0.184 |
20/0082_ST2          |  3.085 | -2.141  0.678  0.482 |  0.288  0.047  0.009 |
20/0118_ST1          |  2.926 | -2.257  0.753  0.595 |  0.064  0.002  0.000 |
20/0149_SQ1          |  2.628 | -1.834  0.498  0.487 | -1.040  0.614  0.157 |
                       Dim.3    ctr   cos2
18/0044_ST2            0.647  0.259  0.024 |
18/0166_SL1           -1.186  0.868  0.317 |
19/0184_ST1           -0.670  0.277  0.082 |
19/0201_ST2           -1.731  1.850  0.173 |
20/0046_NonAdmis2CS   -0.996  0.612  0.066 |
20/0048_ST1            0.372  0.086  0.042 |
20/0057_SD1           -0.443  0.121  0.058 |
20/0082_ST2           -0.492  0.149  0.025 |
20/0118_ST1            1.540  1.465  0.277 |
20/0149_SQ1            1.214  0.909  0.213 |
```

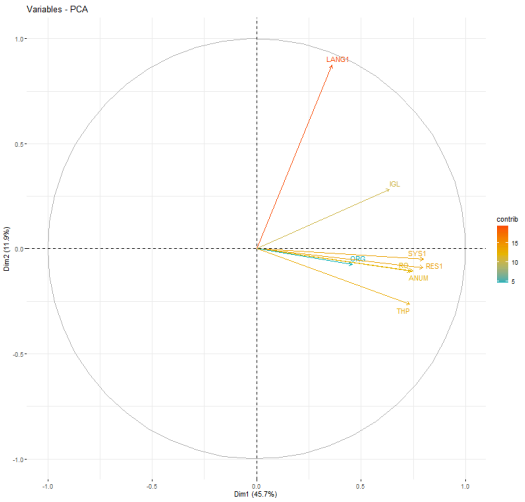
```
Variables
                       Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3
SYS1                 |  0.799 17.473  0.639 | -0.051  0.269  0.003 | -0.042
RES1                 |  0.797 17.370  0.635 | -0.089  0.834  0.008 | -0.119
ANUM                 |  0.747 15.254  0.557 | -0.106  1.170  0.011 | -0.080
RO                   |  0.737 14.859  0.543 | -0.105  1.162  0.011 | -0.143
ORG                  |  0.458  5.732  0.209 | -0.075  0.587  0.006 |  0.828
LANG1                |  0.360  3.553  0.130 |  0.874 80.307  0.765 | -0.130
IGL                  |  0.635 11.052  0.404 |  0.283  8.401  0.080 |  0.245
THP                  |  0.733 14.707  0.537 | -0.263  7.271  0.069 | -0.265
                       ctr   cos2
SYS1                 0.206  0.002 |
RES1                 1.625  0.014 |
ANUM                 0.733  0.006 |
RO                   2.320  0.020 |
ORG                 78.309  0.686 |
LANG1                1.921  0.017 |
IGL                  6.866  0.060 |
THP                  8.020  0.070 |
>
```

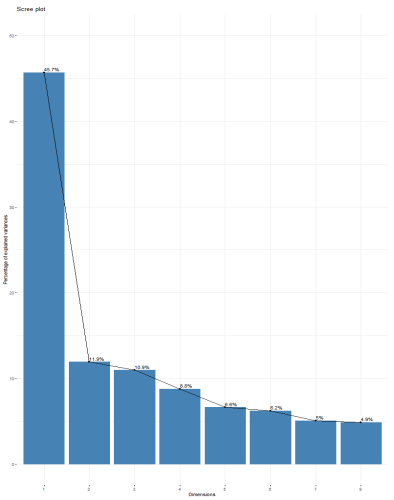
5. Analyse des résultats

5.1. Projection des individus et des variables

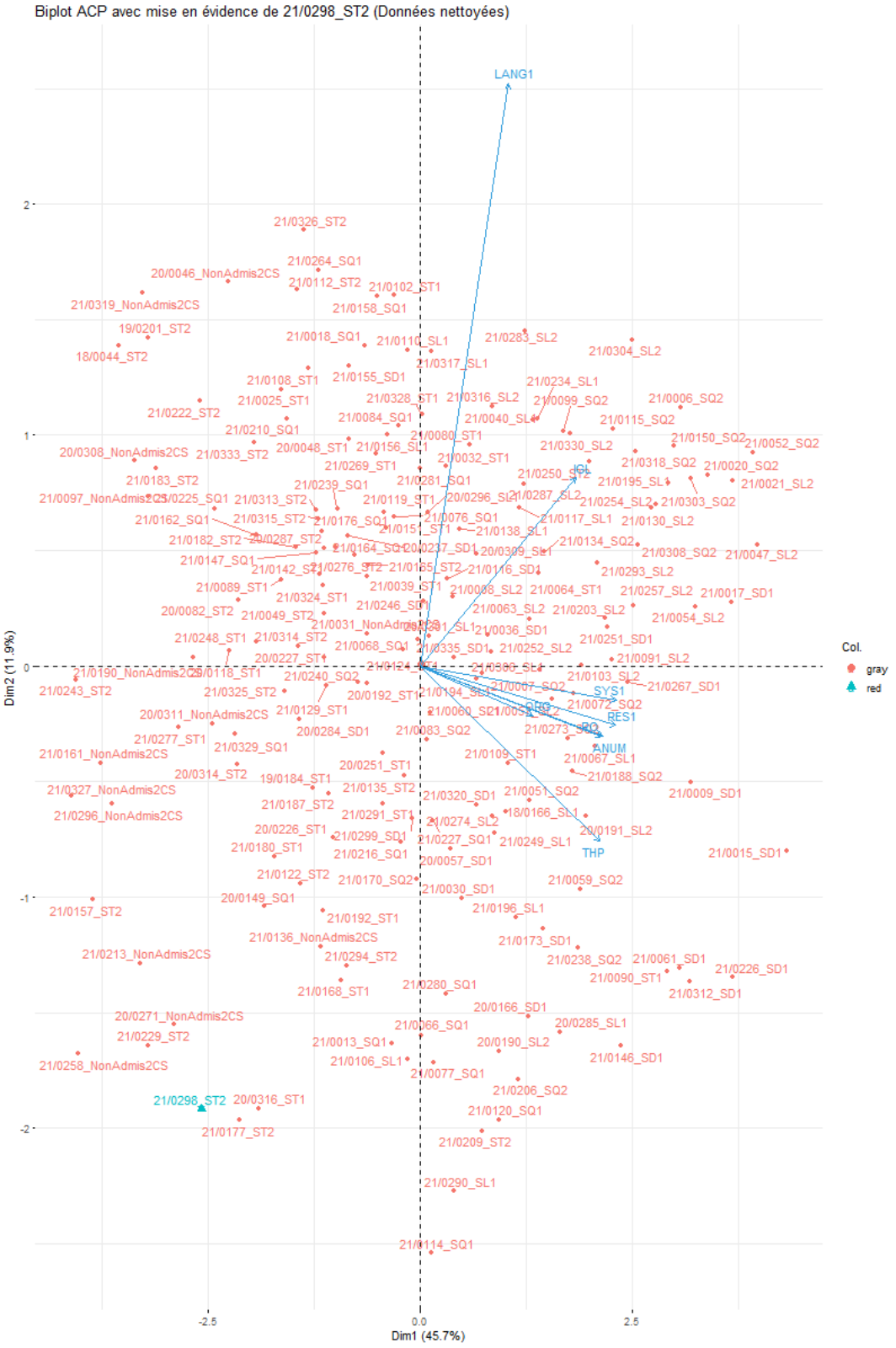
Suite à l'application de l'ACP sur les données nettoyées, nous pouvons observer le positionnement des individus dans l'espace des deux premières composantes principales. Le résumé des résultats montre que la première composante (Dim.1) explique 45.68% de la variance, tandis que la deuxième composante (Dim.2) en explique 11.90%. Ensemble, ces deux premières composantes couvrent près de 57.58% de la variance totale des données. Le biplot issu de cette analyse permet de visualiser les individus par rapport à ces composantes principales, et de détecter ceux qui se regroupent ou qui présentent des valeurs atypiques.



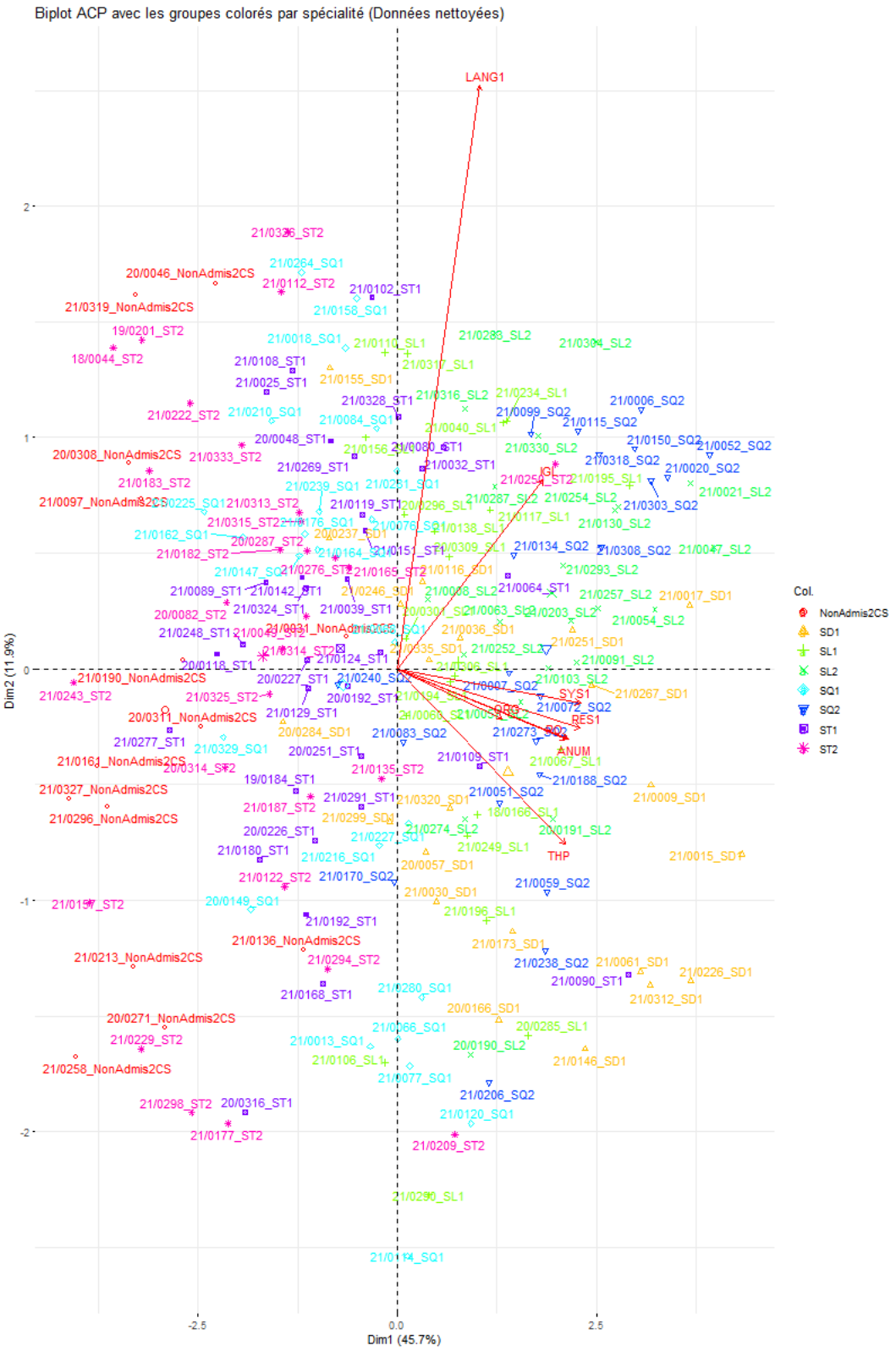




5.2. Mise en évidence ma projection dans biplot


Biplot ACP avec mise en évidence de 21/0298_ST2 (Données nettoyées)
LANG1
IGL
SYS1
RES1
ANUM
THP
21/0298_ST2
Dim2 (11.9%)
Dim1 (45.7%)
Col.
gray
red

5.3. Visualisation par spécialité



Conclusion Générale

L’analyse réalisée sur les données du fichier **1CS S1** a permis de mettre en évidence certaines relations entre les performances des étudiants et leurs spécialités attribuées. Toutefois, les résultats montrent que cette dataset limitée, ne prenant en compte que les notes du premier semestre de la première année (1CS S1), ne permet pas de visualiser de manière suffisamment complète les compétences de chaque individu ni de relier de manière précise ces compétences à leur spécialité.

En effet, les performances des étudiants sur une seule période, sans tenir compte de l'évolution de leurs compétences au fil de l'année, ne fournissent qu'une vision partielle de leur profil académique. De plus, les spécialités attribuées dépendent de multiples facteurs, dont les performances sur l'ensemble de l'année scolaire et pas seulement sur un seul semestre.

Pour obtenir une évaluation plus précise des compétences des étudiants et mieux comprendre leur adéquation avec leur spécialité, il serait nécessaire d'intégrer l'ensemble de leurs notes, y compris celles des semestres suivants. Cette approche plus complète permettrait d'effectuer une analyse plus précise et plus significative des relations entre les performances des étudiants et leurs spécialités. En somme, l'utilisation de toutes les données de la première année jusqu'à la fin du dernier semestre de 1CS serait essentielle pour une visualisation plus fidèle et une interprétation plus complète des compétences individuelles et des regroupements par spécialité.