IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Projet Statistiques

Groupe 8

DATA SCIENCE : PERCEPTION ET UTILISATION PAR LES ÉTUDIANTS DE L'IMT ATLANTIQUE

Présenté par :

BENCHAABANE, BOUAIDA,
EL YOUNSI, HARIM,
HARTMANN, MORSLI,
TAKHCHI

Sous la tutelle de :

Romain Billot
Gilles Coppin
Bernard Gourvennec

SOMMAIRE

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Introduction

Introduction

Contexte

Data Science :
Phénomène de mode

Objectif

Perception et pratique
de la Data Science

Méthode

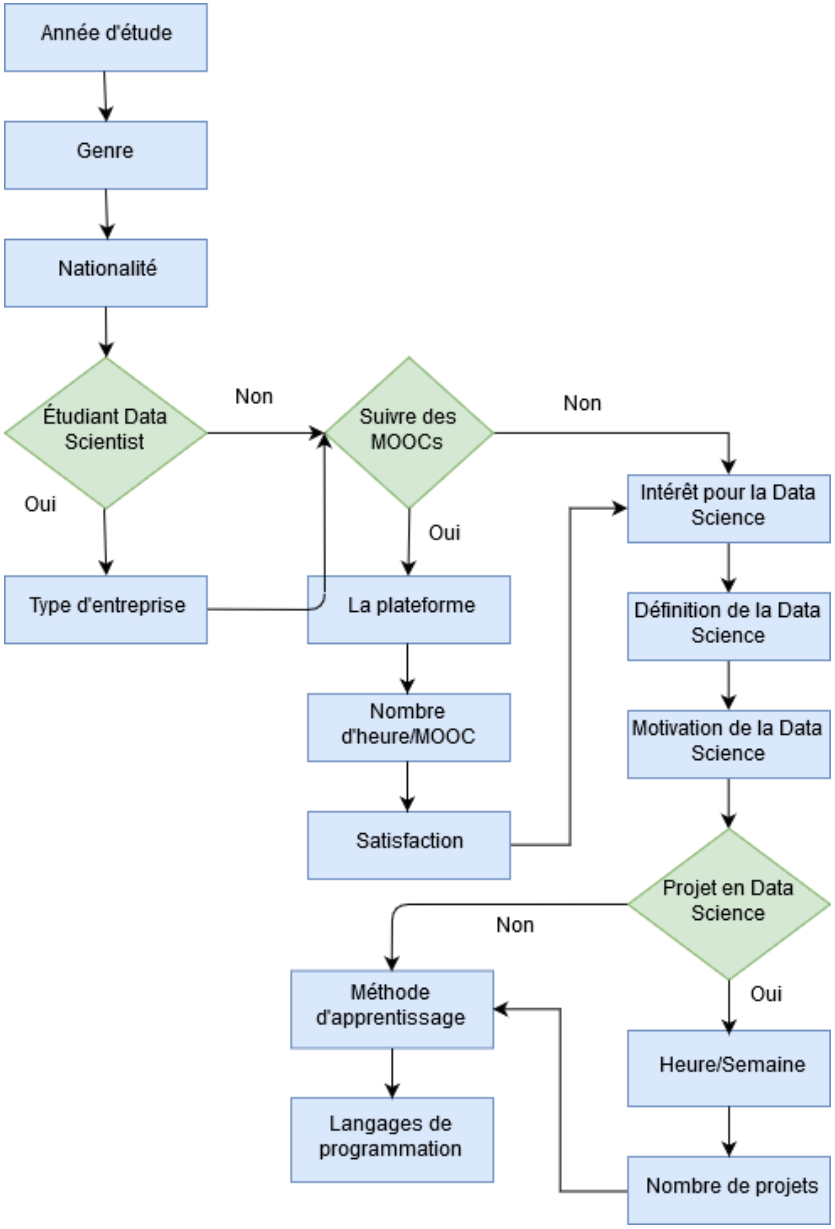
Tests d'hypothèses :
Analyse statistique

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

I-Échantillonnage et Conception du questionnaire

I-Échantillonnage et Conception du questionnaire

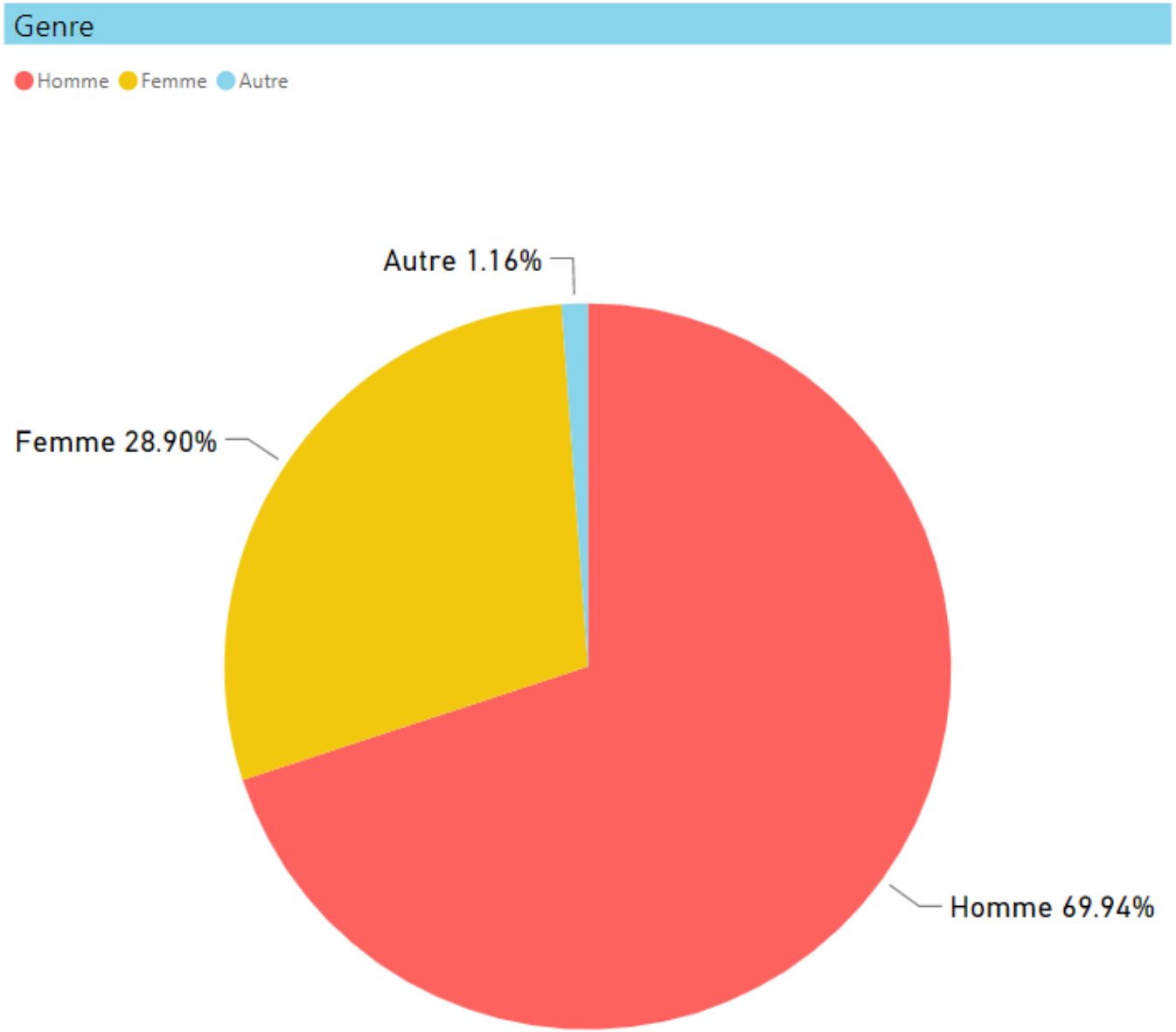
- Identifier la problématique
- Identifier la population
- Formuler des Hypothèses
- Concevoir le Questionnaire



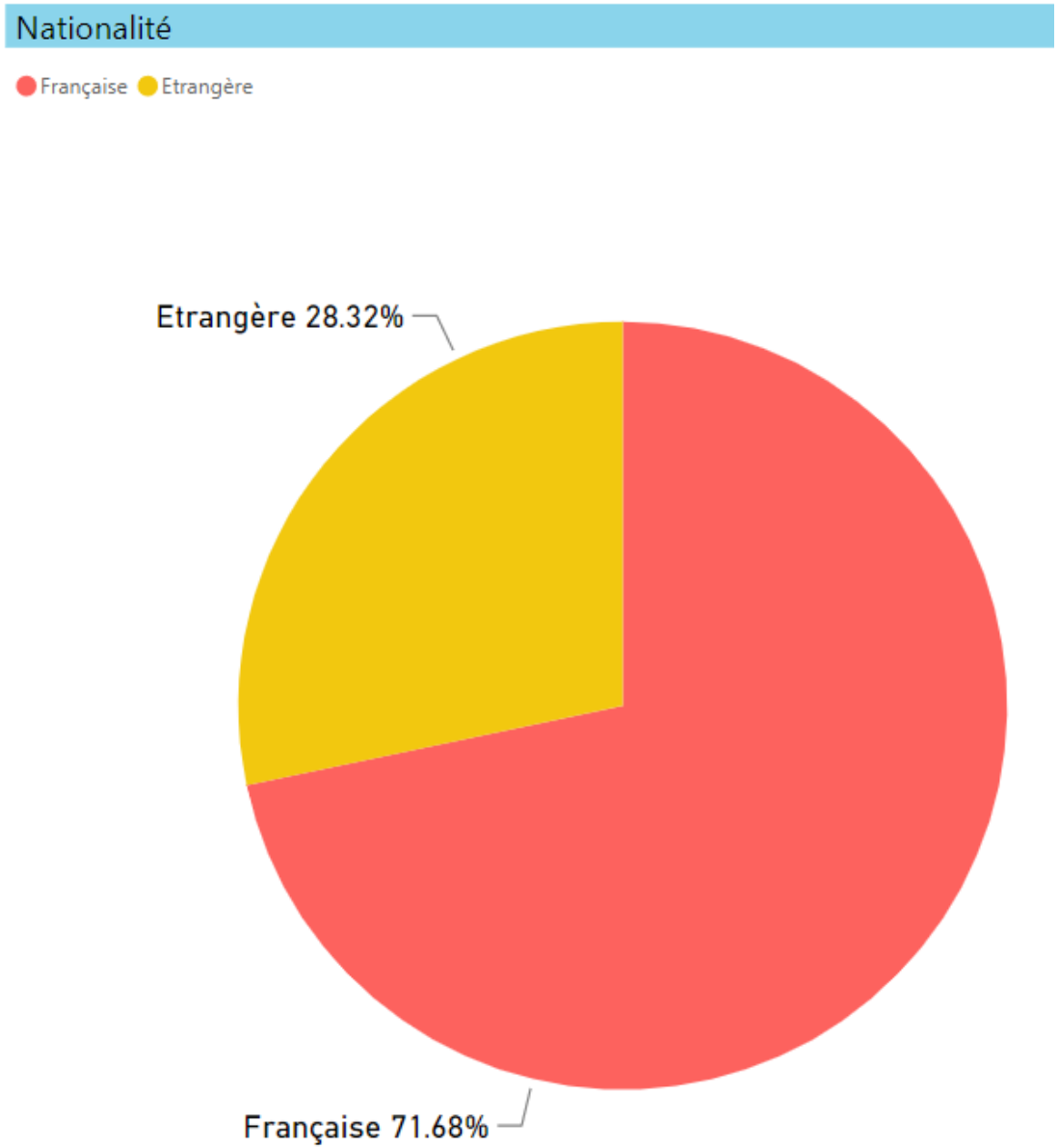
IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

II-Représentativité

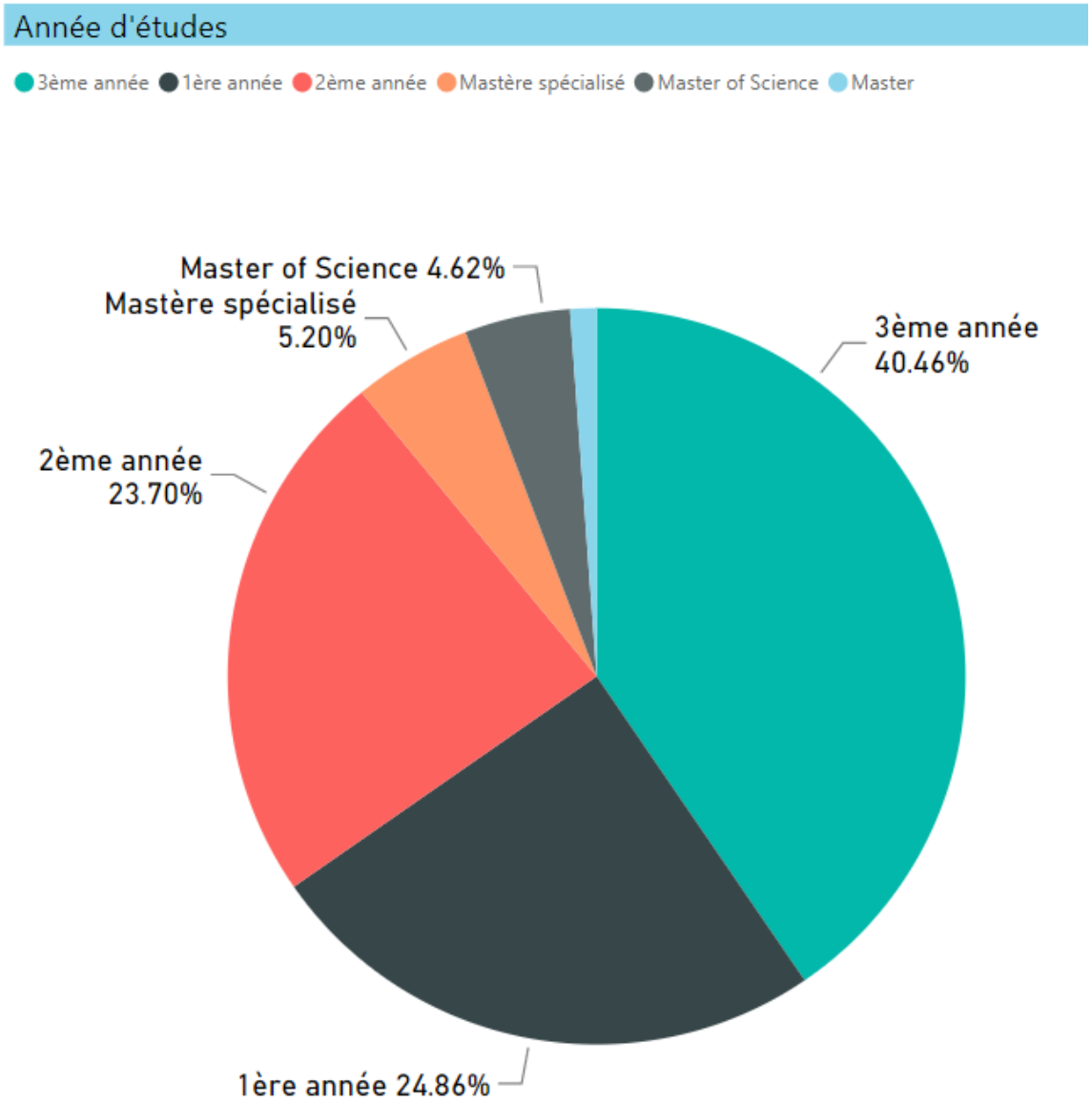
II-Représentativité (|)



II-Représentativité (||)



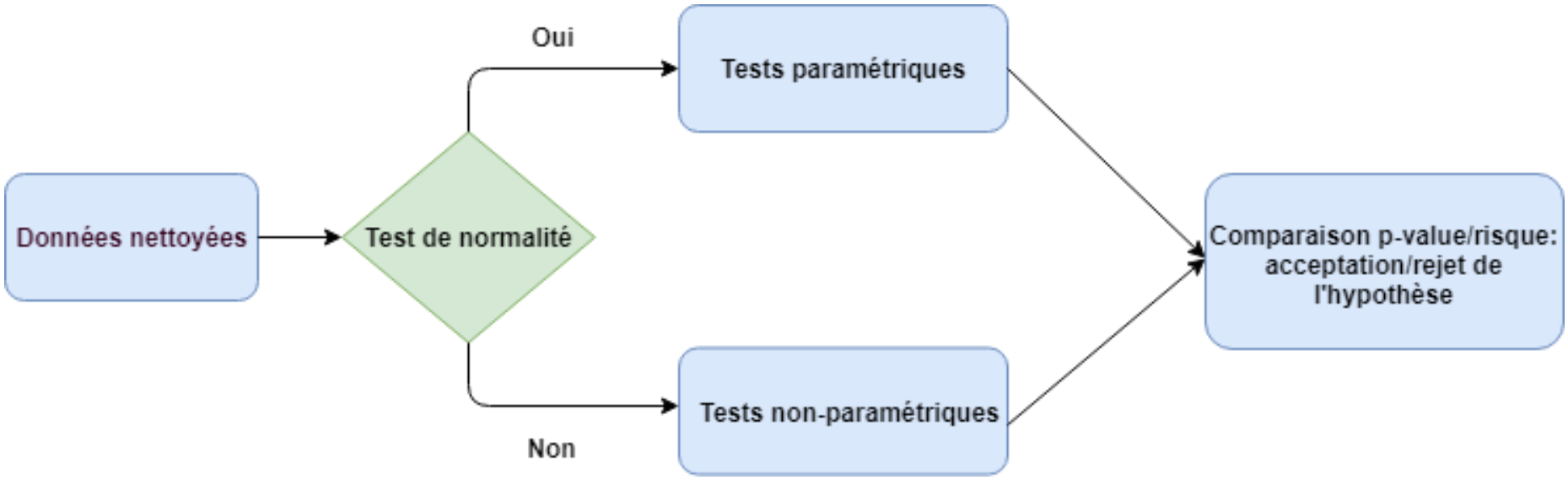
II-Représentativité (III)



IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

III-Tests d'hypothèses

1. Méthodologie de travail



2. Test d'intérêt (|)

Hypothèse nulle

- Le taux de français intéressés par le domaine de la data science est inférieur à celui des étrangers.

Grandeurs statistiques

	Moyenne	Ecart-type
Français	3.73	0.87
Etrangers	4.0	0.88

Test de normalité sur les deux échantillons

Français :

shapiro.test(fr_int)

 # Shapiro-Wilk normality test
 # data: etr_int
 # W = 0.72681, p-value = 3.247e-08

Etrangers :

shapiro.test(etr_int)

 # Shapiro-Wilk normality test
 # data: etr_int
 # W = 0.72681, p-value = 3.247e-08

α = 10%, nous rejetons la normalité pour l'échantillon des français et des étrangers.

2. Test d'intérêt (III)

Test de moyenne

Normalité non vérifiée → ***test non-paramétrique de Wilcoxon***

```
wilcox.test(fr_int,etr_int,alternative = 'g')

# Wilcoxon rank sum test with continuity correction
# data: fr_int and etr_int
# W = 2427, p-value = 0.9878
```

→ ***α = 10%, nous acceptons H0 : Les étrangers sont plus intéressés par la data science que les français.***

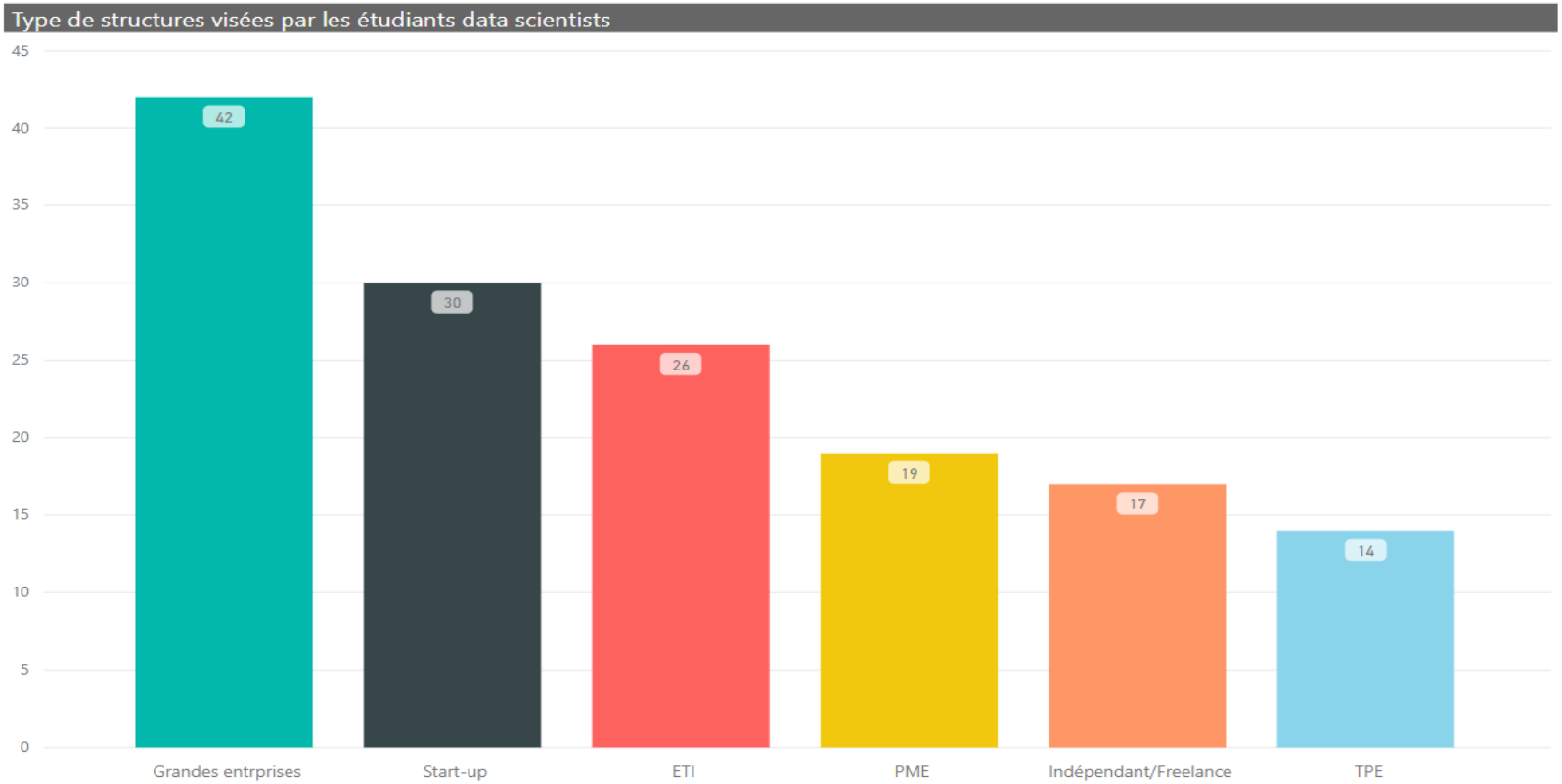
3. Test du χ^2

Hypothèse nulle

- La préférence du type de structures est uniforme pour les étudiants Data scientists.

Test du X2 à 5 degrés de libertés.
p-value = 0.0005963

➡ ***α = 10%, nous rejetons H0 d'équirépartition de la préférence des structures de travail pour les élèves Data Scientists.***



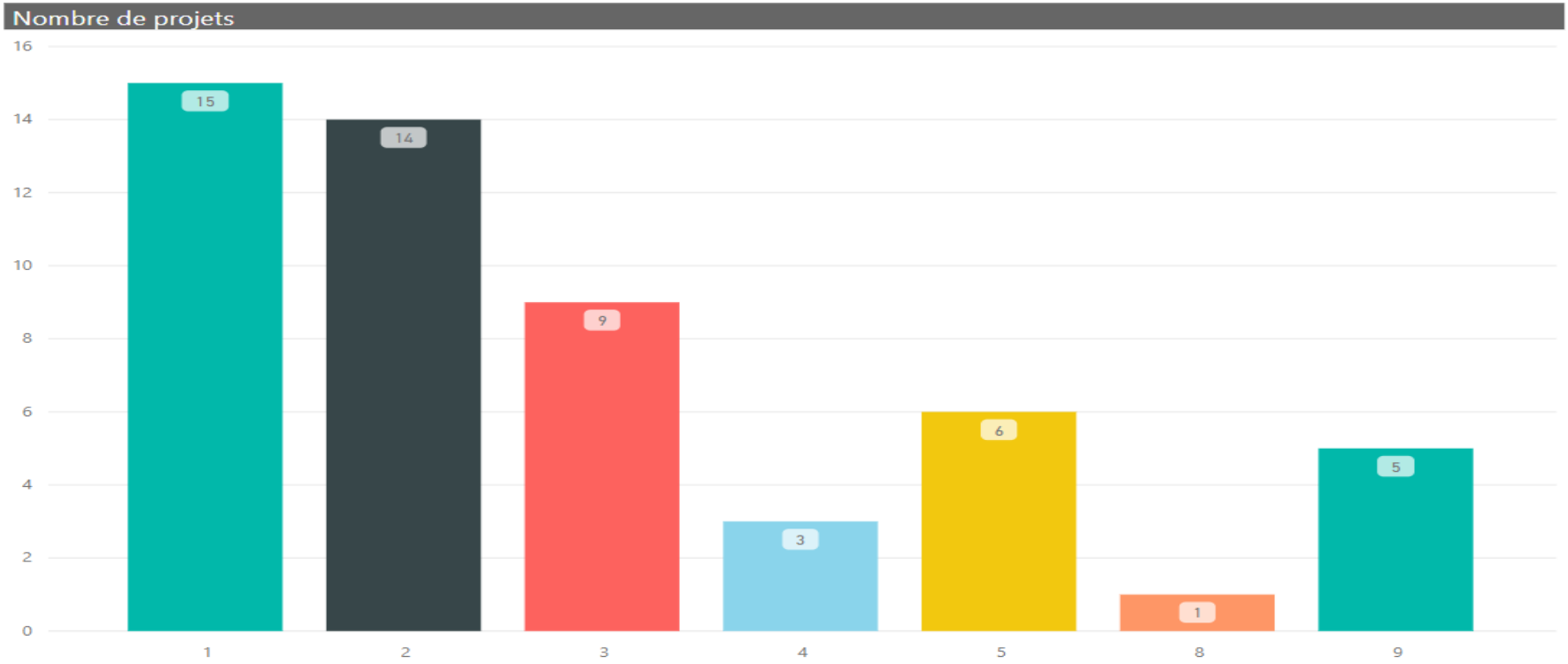
4. Test de Student

Hypothèse nulle

- Un élève ingénieur (toutes filières confondues) travaille en moyenne sur trois projets Data durant sa formation.

t = 0.56743, df = 52, p-value = 0.5729
95 percent confidence interval:
2.473578 3.941516

α = 5%, nous acceptons H0. Un élève ingénieur travaille en moyenne sur trois projets Data Science durant sa formation.



5. Analyse en Composantes Principales (I)

La répartition de la préférence des langages de programmation utilisés en Data Science est uniforme

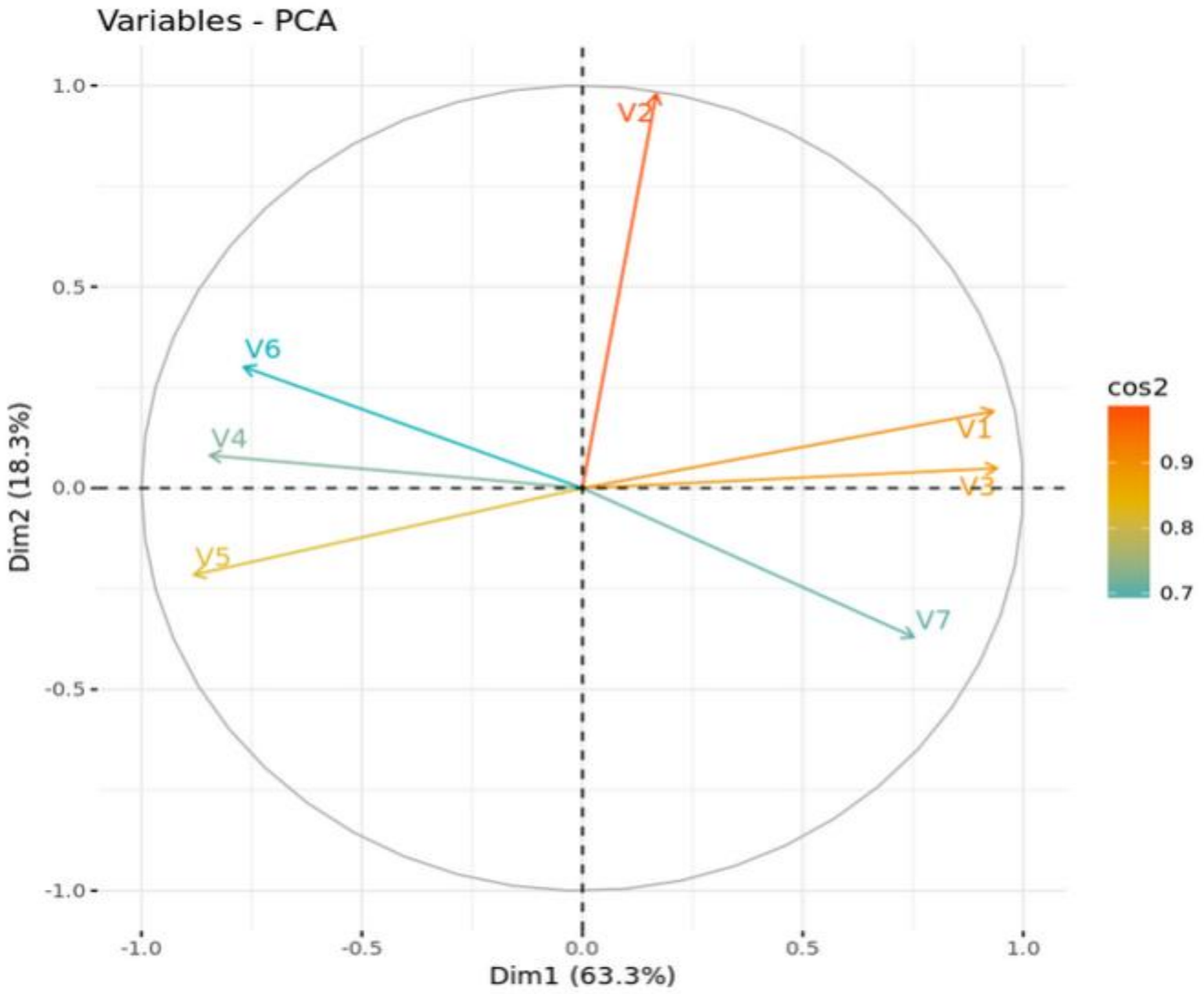
	Français	Etranger	Homme	Femme	1A	2A	3A	MS	MOS	M
Langage										
C	0.258065	0.0816327	0.214876	0.192308	0.470588	0.243902	0.1	0	0.125	1
R	0.33871	0.612245	0.421488	0.384615	0.264706	0.243902	0.571429	0.888889	0.375	1
java	0.306452	0.265306	0.256198	0.384615	0.617647	0.341463	0.128571	0.111111	0.5	1
matlab	0.16129	0.204082	0.157025	0.173077	0.235294	0.121951	0.185714	0.333333	0.125	0
python	0.693548	0.897959	0.752066	0.711538	0.794118	0.731707	0.842857	0.777778	0.75	0.5
scala	0.0483871	0.0816327	0.0330579	0.0961538	0.0588235	0.0243902	0.0857143	0.111111	0	0
sql	0.403226	0.163265	0.280992	0.442308	0.558824	0.365854	0.257143	0.222222	0.375	0.5

Tableau représentant les langages utilisés en Data Science par étudiant, effectif normalisé par catégorie

MS : Mastère spécialisé
M : Master
MOS : Master Of Science

5. Analyse en Composantes Principales (II)

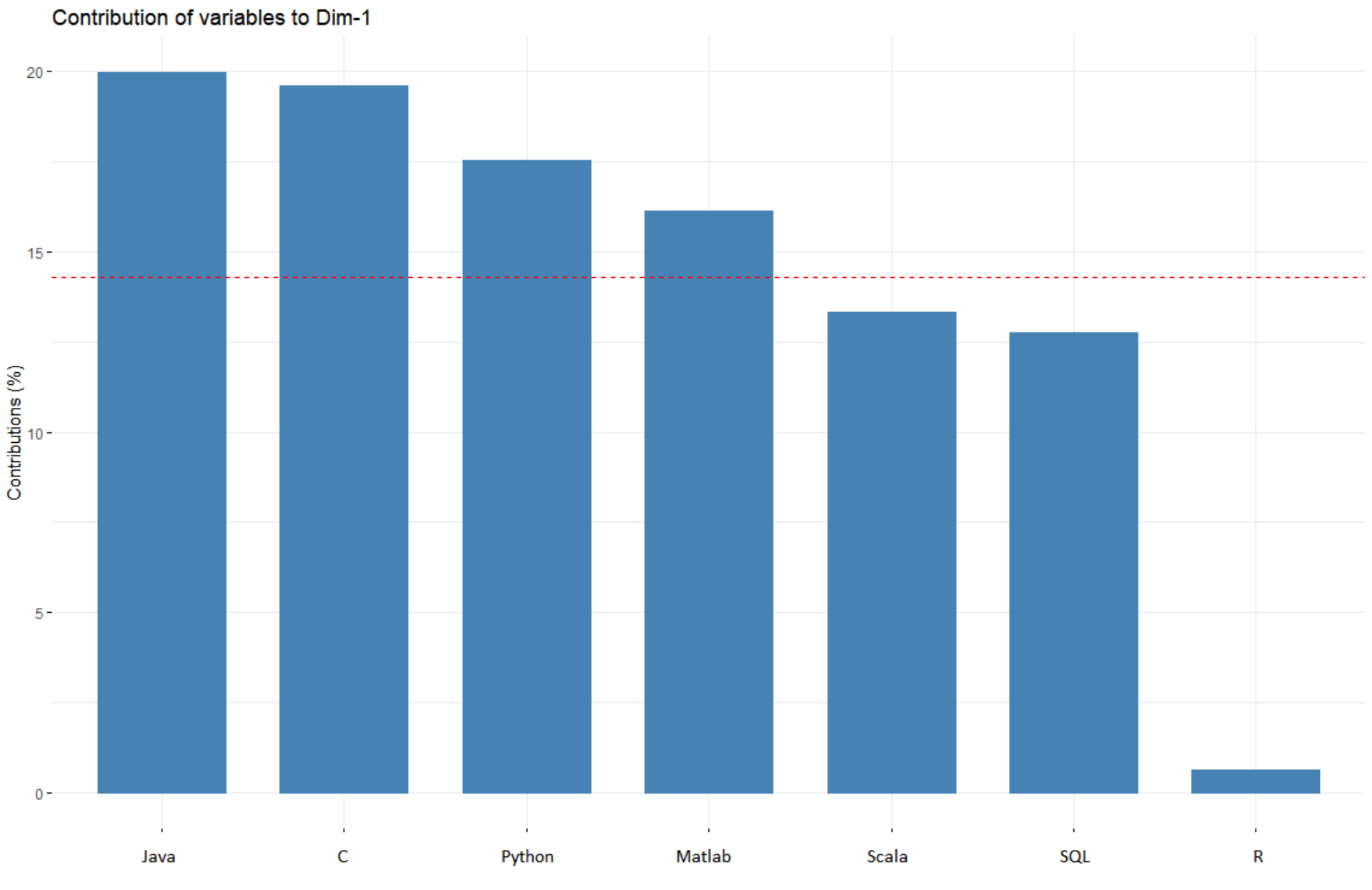
V1 : C
V2 : R
V3 : Java
V4 : Matlab
V5 : Python
V6 : Scala
V7 : SQL



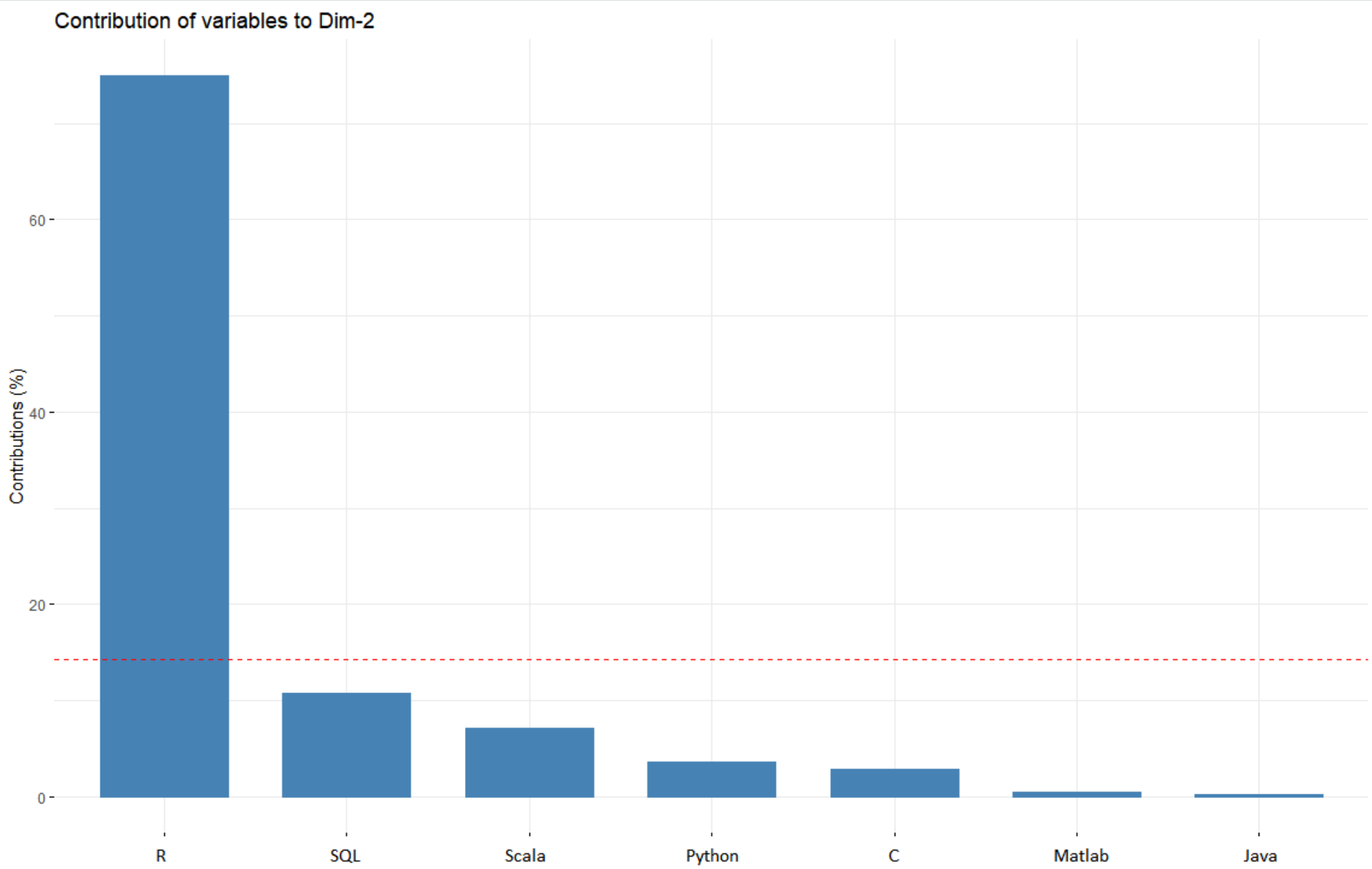
Cercle de corrélation

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

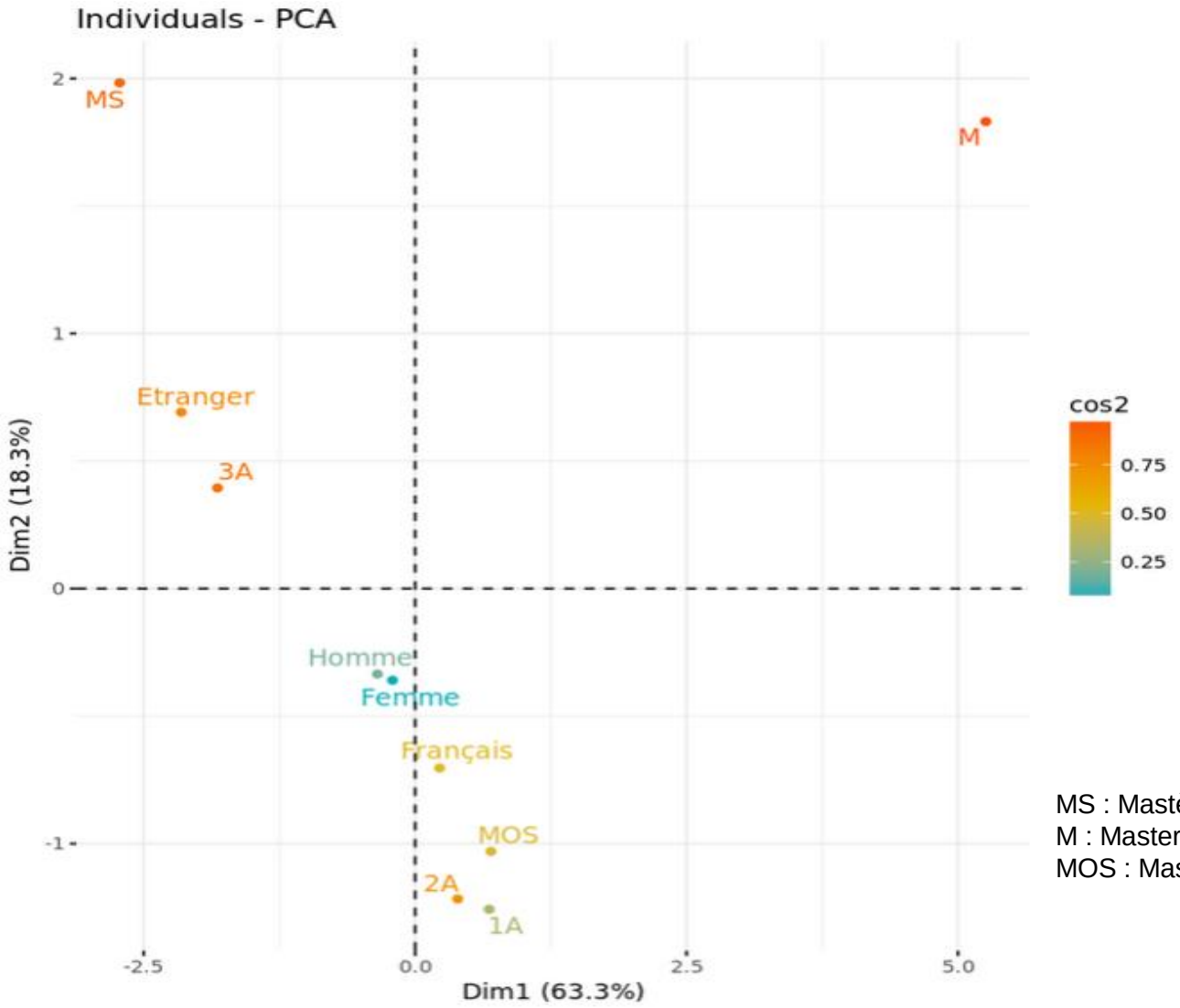
5. Analyse en Composantes Principales (III)



5. Analyse en Composantes Principales (**IV**)



5. Analyse en Composantes Principales (**V**)



MS : Mastère spécialisé
M : Master
MOS : Master Of Science

Représentation des catégories de populations dans l'espace réduit (dim1,dim2)

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

IV-Analyse Critique

IV-Analyse Critique

Données biaisées

- Sur-représentativité de certaines populations.
- Populations d'élèves pas assez expérimentées dans le domaine.
- Questionnaire adressé à des élèves de l'IMT Atlantique

Pour pallier à ces biais, nous proposons les démarches suivantes

- Utiliser un sondage par quota
- Envoyer les questions à des groupes d'écoles ou à des universités représentatives de l'enseignement supérieur français.
- Améliorer la stratégie de communication

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

CONCLUSION

Conclusion

1 Nationalité - Genre

- Les étrangers sont plus intéressés par la data science que les français.
- Les femmes et les hommes sont tout autant intéressés par la Data Science.

2 Motivation - Notoriété

- La Data Science VS électronique / mathématiques / informatique.
- Travailler en Data Science est un choix motivé par l'enjeu des données massives.

3 Préférences

- Les étudiants de l'école apprécient bien les MOOCs de Data Science.
- Python et R restent les deux langages de prédilection des étudiants Data Scientists.

Bibliographie

- XLStat [en ligne]. 2018 (consulté le 12 Novembre 2018).
- Statistical Tools For High-Throughput Data Analysis [en ligne] (consulté le 15 Novembre 2018)
- Google [en ligne] (consulté en ligne le 12 Novembre 2018)

Merci pour votre attention

Avez-vous des questions ?

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom