

ECOLE COMMUNAUTAIRE
DE L'ENSEIGNEMENT
SUPÉRIEUR



Projet

ANALYSE ET
TRAITEMENT DES
DONNÉES

Présenté par :
NGANGA Georgia Floyd Eureka

Supervisé par :
M. LOUKOUBAMA Princy Kénan
Ingénieur statisticien

Table des matières

I.	Introduction.....	4
1.	Contexte	4
2.	Problématique.....	4
3.	Objectifs	4
II.	Cadre expérimental.....	5
III.	Partie 1.....	5
1.	Description du jeu de données	5
2.	Création de nouvelles variables	6
3.	Statistiques.....	7
a)	Répartition des candidats par sexe suivant les pays	7
b)	Répartition des candidats admis par pays	9
c)	Pourcentage des candidats admis suivant les pays	11
d)	Tableau représentant les 5 premiers candidats admis par pays	12
IV.	Partie 2	15
1.	Vérification de l'assertion ” les pays de l'Afrique de l'Ouest ont des meilleures performances au Bac que les pays de l'Afrique Centrale.”	15
	Création de la colonne " Mention au Bac "	15
	Analyse Factorielle des Correspondances (AFC).....	16
	Test de khi-deux	17
2.	La performance des candidats au Bac est-elle liée au pays ?	18
3.	Lien entre la performance au Bac et la région.....	23
4.	Lien entre la région et l'admission au concours.....	24
6.	Lien entre l'ancienneté du bac et l'admission au concours.....	27
7.	Modélisation par Machine Learning	29
8.	Vérification de la robustesse du modèle.....	31
	La courbe ROC et l.....	31
	Test de HosmerLemeshow.....	32
V.	Conclusion	33

Table des illustrations

Graphique 1: Répartition des candidats par sexe selon les pays	9
Graphique 2: Répartition des candidats admis par pays.....	10
Graphique 3: Pourcentage des candidats admis suivant les pays	12
Graphique 4: AFC Mention au bac vs R	
Tableau 1: Effectifs des candidats par sexe et par pays	8
Tableau 2: Répartition des candidats admis par pays	10
Tableau 3: Répartition des candidats admis et échoués par pays en pourcentage (%)	11
Tableau 4: Tableau représentant les 5 premiers candidats admis par pays.....	13
Tableau 5: Région x Mention au Bac	16
Tableau 6: Mention au Bac vs Pays	20
Tableau 7: Mention au Bac vs Région	23
Tableau 8: Région vs Admission au concours.....	24
égin.....	16
Graphique 5: Répartition de la moyenne au Bac par pays.....	18
Graphique 6: Répartition des candidats admis et échoués selon leur région d'origine	24
Graphique 7: Moyenne au Bac vs Admission au concours	27
Graphique 8: Ancienneté du bac vs admission au concours	28
Graphique 9: Répartition des candidats admis et échoués au concours	30
Graphique 10: Courbe Roc.....	31

I. Introduction

1. Contexte

L'École Africaine de la Météorologie et de l'Aviation Civile (EAMAC), située à Niamey au Niger, joue un rôle stratégique dans la formation des professionnels de la météorologie et de l'aviation civile en Afrique. Chaque année, cette institution accueille des candidats provenant de plusieurs pays africains à travers un concours d'admission rigoureux. Ce processus de sélection, basé sur quatre matières principales (mathématiques, physique, anglais et français), constitue la porte d'entrée vers une formation spécialisée répondant aux besoins du secteur aéronautique et météorologique du continent.

L'analyse des données d'admission revêt une importance particulière pour comprendre les dynamiques de recrutement et identifier les facteurs qui influencent la réussite des candidats. Dans un contexte où l'équité géographique et la diversité des profils constituent des enjeux majeurs, il devient essentiel d'examiner objectivement les résultats du concours 2025.

2. Problématique

L'analyse des données d'admission au concours EAMAC 2025 révèle des enjeux cruciaux pour l'équité et l'efficacité du processus de sélection. Face à la diversité des profils candidats provenant de différents pays africains, une question fondamentale se pose : quels sont les véritables déterminants de la réussite au concours ?

Plusieurs hypothèses méritent d'être explorées scientifiquement. D'une part, l'origine géographique des candidats pourrait-elle créer des inégalités de chances ? Certaines régions africaines bénéficient-elles d'avantages systémiques en termes de qualité de formation ou de préparation aux concours ? D'autre part, les caractéristiques individuelles des candidats - leur maturité académique, leur expérience post-baccalauréat, ou leur genre - influencent-elles significativement leurs performances ?

Un élément particulièrement intriguant émerge des conclusions d'une précédente étude de 2024, qui pointait une supériorité des performances au baccalauréat chez les candidats d'Afrique de l'Ouest par rapport à ceux d'Afrique Centrale. Cette observation soulève une interrogation majeure : cette tendance se traduit-elle effectivement par un avantage concurrentiel lors du concours EAMAC, ou s'agit-il d'un préjugé à déconstruire par une analyse rigoureuse des données ?

3. Objectifs

Cette analyse vise à fournir à la Direction générale de l'EAMAC une évaluation complète et objective des résultats du concours 2025. Les objectifs spécifiques sont les suivants :

Objectif principal : Analyser les facteurs déterminants de l'admission au concours EAMAC 2025 et identifier les tendances significatives dans les résultats.

Objectifs spécifiques :

- Caractériser le profil démographique et géographique des candidats
- Évaluer les disparités de performance entre les pays et régions d'origine
- Examiner l'influence des facteurs individuels (âge, ancienneté du baccalauréat, sexe) sur les résultats
- Vérifier statistiquement les hypothèses concernant les différences régionales de performance
- Développer un modèle prédictif pour optimiser le processus de sélection future

Cette approche méthodologique permettra d'éclairer les décisions stratégiques de l'institution et de contribuer à l'amélioration continue du processus d'admission, dans un souci d'équité et d'excellence académique.

II. Cadre expérimental

Dans le cadre de cette étude, nous nous intéressons à l'analyse des résultats du concours d'entrée à l'EAMAC pour l'année 2025. Celle-ci se déroulera en **deux grandes parties**.

La **première partie** portera sur l'analyse descriptive et exploratoire des données issues du concours, avec la création de nouvelles variables (âge, pays, moyenne pondérée, rang, décision, etc.) et la production de tableaux et graphiques permettant de mieux comprendre les caractéristiques des candidats et les tendances d'admission.

La **seconde partie** visera à évaluer des relations statistiques entre certaines variables clés (région, pays, ancienneté du Bac, etc.) et la performance ou l'admission des candidats. Elle se prolongera par la mise en œuvre d'un **modèle de classification en machine learning**, destiné à prédire l'admission au concours sur la base de variables explicatives pertinentes.

Toutes les étapes de traitement, d'analyse et de modélisation seront **entièvement réalisées en langage Python**.

III. Partie 1

1. Description du jeu de données

La présente étude s'appuie sur la base de données intitulée “**EAMAC.xlsx**”, un fichier Excel recueilli dans le cadre du concours d'admission à l'École Africaine de la Météorologie et de l'Aviation Civile (EAMAC) pour l'année 2025. Ce fichier constitue

le socle empirique de notre analyse et renferme des informations précieuses sur les profils et performances académiques des candidats.

La base comporte un total de **21 000 observations**, représentant chacun un candidat au concours. Ces observations sont structurées autour de **8 variables initiales**, que nous présentons ci-après.

- **Matricule** : Il s'agit d'un identifiant unique attribué à chaque candidat. Plus qu'un simple code, le matricule regroupe des informations encodées telles que le **sexe** du candidat (lettre H pour homme, F pour femme) ainsi que sa **date de naissance** (jour, mois, année). Ces données seront exploitées pour dériver des variables comme l'âge ou encore le sexe, nécessaires à l'analyse.
- **Pays** : Cette variable indique le pays d'origine du candidat.
- **Année admission Bac** : Cette variable renseigne sur l'année à laquelle le candidat a obtenu son baccalauréat. Elle est essentielle pour mesurer l'ancienneté du diplôme et évaluer la maturité académique.
- **Moyenne au Bac** : Elle correspond à la moyenne générale obtenue par le candidat lors de son examen de baccalauréat. Elle constitue une variable clé dans l'étude des performances scolaires pré-concours.
- **Maths, Physique, Français, Anglais** : Ces quatre variables correspondent aux notes obtenues par chaque candidat aux différentes épreuves du concours d'entrée. Chaque matière est associée à un coefficient de pondération spécifique, ce qui permettra le calcul d'une **moyenne pondérée d'admission**, essentielle pour déterminer les résultats définitifs.

Ainsi, ce jeu de données offre une base solide pour conduire une analyse approfondie des performances des candidats et des facteurs susceptibles d'influencer leur admission à l'EAMAC. Il permettra également d'explorer des relations statistiques entre plusieurs variables explicatives (âge, pays, région, ancienneté du Bac, etc.) et le résultat au concours.

2. Création de nouvelles variables

Dans le but d'enrichir notre analyse et de tirer pleinement parti des informations encodées dans la base, nous avons procédé à la création de plusieurs variables dérivées, toutes essentielles à l'interprétation des résultats du concours.

Nous avons tout d'abord généré une colonne nommée **Ndate**, correspondant à la **date de naissance complète** de chaque candidat. Celle-ci a été extraite directement à partir du **matricule**, en identifiant successivement le jour, le mois et l'année de naissance, encodés après l'indication du sexe dans le code. La nouvelle

variable est ensuite formatée selon la structure classique **jj/mm/aaaa**, facilitant son exploitation dans le calcul de l'âge ou d'autres analyses temporelles.

À partir de la date de naissance des candidats, une seconde variable a été construite : **l'âge**, calculé en prenant **l'année 2025 comme référence**, année du concours. L'âge des candidats se situe ainsi entre **16 et 25 ans**, ce qui correspond à la tranche d'âge attendue pour ce type de sélection.

Par ailleurs, une colonne "Moyenne" a été ajoutée pour représenter les performances des candidats au concours. Cette moyenne a été calculée en tenant compte de l'importance relative accordée à chaque matière dans le barème officiel. Elle est ensuite arrondie à deux décimales pour faciliter la lecture. Les moyennes obtenues varient entre **8,95** et **18,62**. Ce qui traduit une grande disparité de niveau entre les candidats..

Pour compléter cette évaluation, une variable **Décision** a été introduite. Elle détermine si le candidat est **admis** (lorsque la moyenne est supérieure ou égale à 16) ou **échoué** (dans le cas contraire).

Enfin, nous avons généré une colonne **Rang**, attribuant à chaque candidat une position dans le classement général selon sa performance. Une attention particulière a été portée à la forme du rang : lorsqu'un homme occupe la première position, le rang affiché est "**1er**", tandis qu'il devient "**1ère**" si c'est une femme. Les autres rangs sont simplement notés avec un "e" final (ex. : 2e, 3e, etc.).

Ces nouvelles variables nous permettront de mener des analyses plus précises, de comparer les performances entre les sous-groupes de candidats et de construire, par la suite, des modèles prédictifs robustes.

3. Statistiques

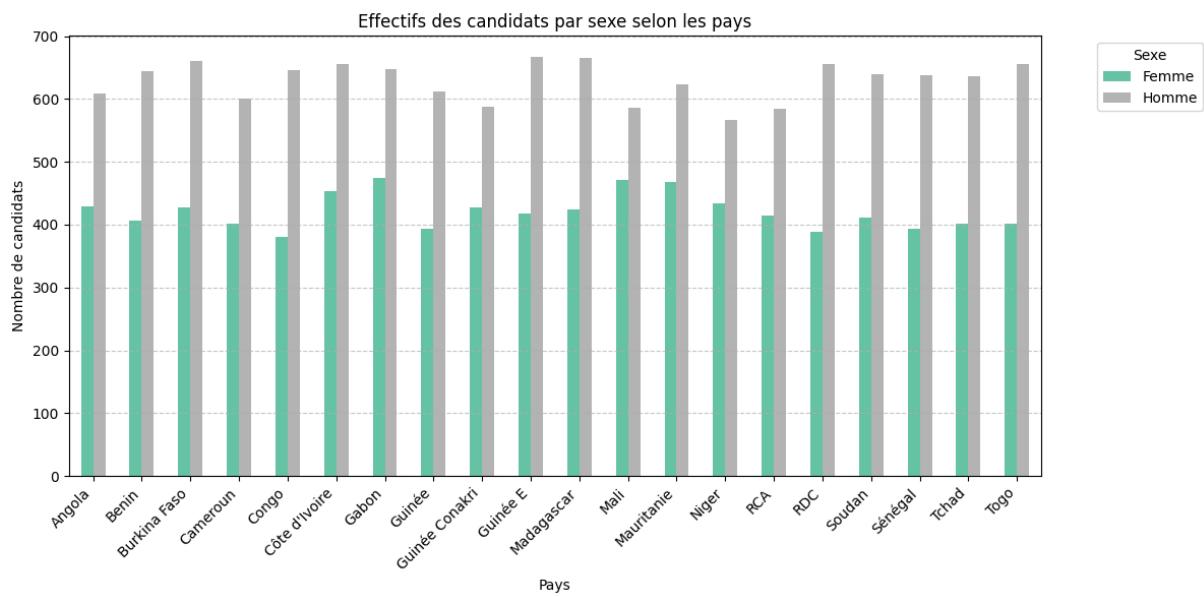
a) Répartition des candidats par sexe suivant les pays

Les données montrent que les hommes sont systématiquement plus nombreux que les femmes à passer le concours EAMAC, avec une moyenne d'environ 6 hommes pour 4 femmes. Le Burkina Faso enregistre le plus grand nombre de candidats, suivi du Mali, de la Mauritanie et du Togo, tandis que la Guinée-Bissau et le Niger en comptent le moins.

Cette différence entre hommes et femmes n'est pas surprenante dans les domaines techniques comme la météorologie et l'aviation civile, qui attirent traditionnellement plus d'hommes. Cela peut s'expliquer par des facteurs culturels et sociaux qui influencent encore les choix d'orientation des jeunes selon leur sexe.

Tableau 1: Effectifs des candidats par sexe et par pays

Pays	Femme	Homme
Angola	429	609
Benin	407	644
Burkina Faso	428	661
Cameroun	402	601
Congo	381	646
Côte d'Ivoire	454	655
Gabon	475	647
Guinée	394	612
Guinée Conakri	427	587
Guinée E	417	667
Madagascar	424	665
Mali	471	586
Mauritanie	468	623
Niger	434	566
RCA	414	585
RDC	389	656
Soudan	411	639
Sénégal	394	637
Tchad	402	636
Togo	401	656



Graphique 1: Répartition des candidats par sexe selon les pays

b) Répartition des candidats admis par pays

Les résultats d'admission au concours montrent des écarts marqués entre les pays, soulignant des dynamiques régionales contrastées. Trois pays se démarquent de façon nette : **la Côte d'Ivoire et le Mali**, qui comptent chacun près de **190 candidats admis**, suivis de **le Sénégal**, avec environ **180 admissions**. Ces chiffres traduisent une forte représentation de ces pays parmi les meilleurs résultats du concours.

Un second groupe de pays affiche des performances honorables, avec des effectifs admis compris entre **150 et 170**. Il s'agit notamment du **Bénin**, du **Cameroun**, de la **Guinée**, du **Niger** et du **Togo**. Ces pays forment un noyau intermédiaire, montrant une bonne préparation générale, bien que légèrement en retrait par rapport aux trois premiers.

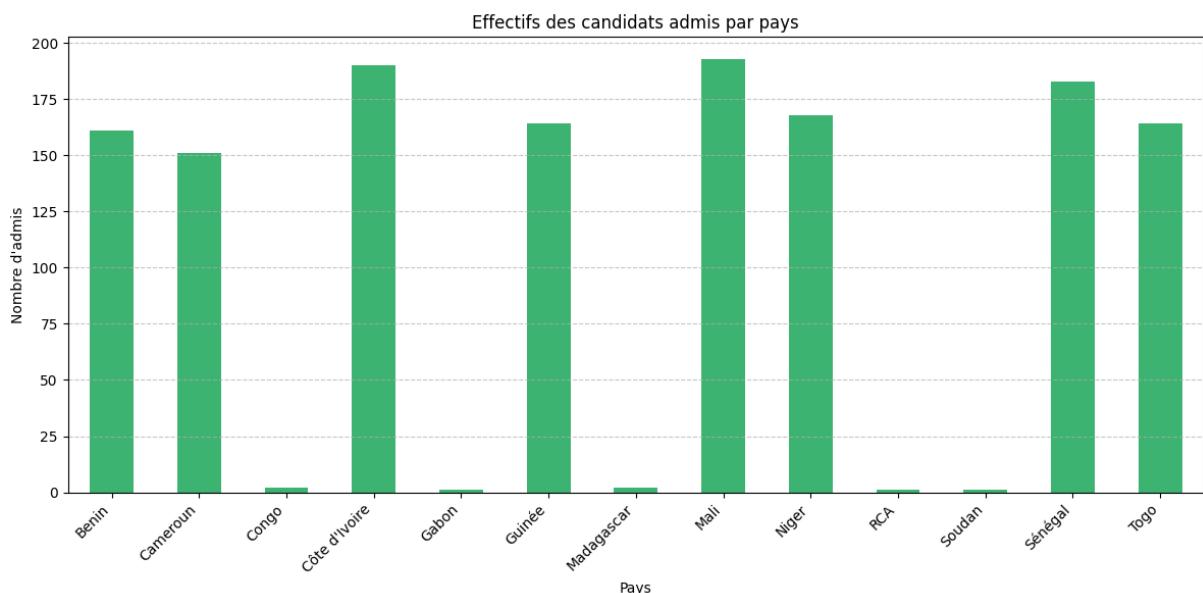
En revanche, certains pays enregistrent des résultats particulièrement faibles. **Le Congo, Madagascar, la République Centrafricaine (RCA) et le Soudan** comptent **moins de 5 admis chacun**, illustrant des difficultés marquées d'accès ou de réussite à ce concours. Le cas du **Gabon** est encore plus préoccupant, avec **aucune admission recensée**, malgré la participation de candidats.

Cette forte disparité entre pays peut s'expliquer par des facteurs variés, allant de la qualité des systèmes éducatifs à l'accès à l'information sur le concours, en passant par la préparation spécifique des candidats.

Tableau 2: Répartition des candidats admis par pays

Pays	Nombre de candidats admis
Benin	161
Cameroun	151
Congo	2
Côte d'Ivoire	190
Gabon	1
Guinée	164
Madagascar	2
Mali	193
Niger	168
RCA	1
Soudan	1
Sénégal	183
Togo	164

Remarque : Les pays absents de ce tableau n'ont enregistré aucun candidat admis au concours.



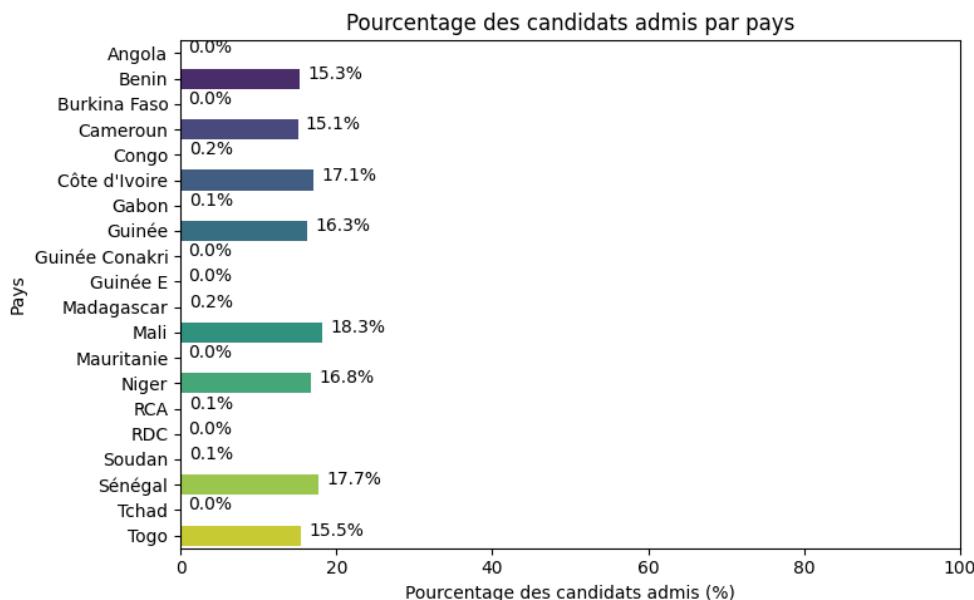
Graphique 2: Répartition des candidats admis par pays

c) Pourcentage des candidats admis suivant les pays

Les taux de réussite au concours varient fortement d'un pays à l'autre. Le **Mali** (18,3 %), le **Sénégal** (17,7 %) et la **Côte d'Ivoire** (17,1 %) enregistrent les meilleurs résultats, confirmant la bonne performance globale des pays d'Afrique de l'Ouest. Un groupe intermédiaire composé du **Niger**, de la **Guinée**, du **Togo**, du **Bénin** et du **Cameroun** affiche des taux entre **15 % et 17 %**. À l'opposé, des pays comme **le Congo, le Gabon, Madagascar, la RCA et le Soudan** ont des taux quasi nuls, tandis que plusieurs autres (**Angola, Burkina Faso, RDC, etc.**) n'ont eu **aucun admis**.

Tableau 3: Répartition des candidats admis et échoués par pays en pourcentage (%)

Pays	Admis	Échoué	Total
Angola	0.0%	100%	100%
Benin	15.3%	84.7%	100%
Burkina Faso	0.0%	100%	100%
Cameroun	15.1%	84.9%	100%
Congo	0.2%	99.8%	100%
Côte d'Ivoire	17.1%	82.9%	100%
Gabon	0.1%	99.9%	100%
Guinée	16.3%	83.7%	100%
Guinée Conakri	0.0%	100%	100%
Guinée E	0.0%	100%	100%
Madagascar	0.2%	99.8%	100%
Mali	18.3%	81.7%	100%
Mauritanie	0.0%	100%	100%
Niger	16.8%	83.2%	100%
RCA	0.1%	99.9%	100%
RDC	0.0%	100%	100%
Soudan	0.1%	99.9%	100%
Sénégal	17.7%	82.3%	100%
Tchad	0.0%	100%	100%
Togo	15.5%	84.5%	100%



Graphique 3: Pourcentage des candidats admis suivant les pays

d) Tableau représentant les 5 premiers candidats admis par pays

L'analyse des 5 meilleurs candidats admis par pays révèle des performances remarquables, en particulier de la part du Cameroun, qui se place à la première du concours avec une moyenne exceptionnelle de 18,62, suivie d'un autre candidat camerounais classé deuxième avec 18,49. Le Mali se distingue également avec une candidate à 18,27, occupant la troisième place nationale. Le Sénégal et le Niger complètent ce peloton de tête avec des moyennes avoisinant les 18,12, confirmant la compétitivité de ces pays.

On observe que les candidats les mieux classés proviennent majoritairement d'Afrique de l'Ouest et du Centre, notamment du Cameroun, du Mali, du Sénégal et du Niger, qui affichent des profils académiques solides. Le Bénin, la Côte d'Ivoire, la Guinée et le Togo présentent également des moyennes très élevées parmi leurs cinq meilleures, souvent supérieures à 17,5, mais légèrement en retrait par rapport aux premières positions.

En revanche, certains pays comme le Congo, le Gabon, la RCA, le Soudan et Madagascar figurent plus bas dans le classement, avec des moyennes autour de 16 chez leurs meilleurs candidats. Cela montre un écart assez net de performance entre les pays.

Il est aussi intéressant de noter que les femmes sont bien représentées parmi les meilleures, avec plusieurs d'entre elles figurant dans les premières places de leurs pays respectifs, voire dans le top national. Cette présence féminine dans les très bonnes moyennes témoigne d'un équilibre croissant dans la réussite entre les sexes.

Dans l'ensemble, ce tableau met en lumière non seulement les écarts de performance entre pays, mais aussi la capacité de certains systèmes éducatifs à préparer des candidats de très haut niveau.

Tableau 4: Tableau représentant les 5 premiers candidats admis par pays

Matricule	Pays	Sexe	Moyenne	Rang
5015Fo8082001	Benin	Femme	17.96	8e
5033Ho8072008	Benin	Homme	17.95	9e
5853Ho1092005	Benin	Homme	17.9	14e
8820Fo1052007	Benin	Femme	17.87	17e
4947Ho6072009	Benin	Homme	17.65	29e
8877Fo9022005	Cameroun	Femme	18.62	1ère
5860Ho9082003	Cameroun	Homme	18.49	2e
6110Ho5032005	Cameroun	Homme	18.04	6e
4955Ho8052003	Cameroun	Homme	17.79	20e
7064H25082001	Cameroun	Homme	17.53	44e
3635Fo4092002	Congo	Femme	16.09	1159e
6889Ho4042005	Congo	Homme	16.0	1360e
4888F10072005	Côte d'Ivoire	Femme	17.93	10e
6610Ho7092000	Côte d'Ivoire	Homme	17.75	23e
2948Fo8072006	Côte d'Ivoire	Femme	17.57	38e
8719F24012009	Côte d'Ivoire	Femme	17.49	49e
4611Ho3072004	Côte d'Ivoire	Homme	17.42	64e
7981Fo2012003	Gabon	Femme	16.19	997e
4789Ho5022008	Guinée	Homme	17.77	21e
3164Ho2092004	Guinée	Homme	17.68	26e
8420Fo2032000	Guinée	Femme	17.56	40e
2055Fo3032009	Guinée	Femme	17.54	42e
4899Fo4052006	Guinée	Femme	17.53	43e

3914Ho1032000	Madagascar	Homme	16.15	1044e
6350F23102009	Madagascar	Femme	16.08	1184e
5687Fo3012007	Mali	Femme	18.27	3e
4984H26052007	Mali	Homme	17.89	15e
4926Ho8052002	Mali	Homme	17.76	22e
2428Ho7072003	Mali	Homme	17.69	25e
5073F19102005	Mali	Femme	17.57	39e
8384Ho3072004	Niger	Homme	18.12	5e
8874Ho6052009	Niger	Homme	17.91	11e
8869Fo1082004	Niger	Femme	17.89	16e
7631Ho8042004	Niger	Homme	17.87	18e
2638F22022008	Niger	Femme	17.66	28e
7264Fo4042005	RCA	Femme	16.24	906e
7281Ho6092000	Soudan	Homme	16.64	436e
8484Ho3012000	Sénégal	Homme	18.12	4e
7522Ho1022000	Sénégal	Homme	17.98	7e
2441Ho8062005	Sénégal	Homme	17.9	13e
7932H11042007	Sénégal	Homme	17.81	19e
5857Fo2062009	Sénégal	Femme	17.67	27e
2841Fo3062004	Togo	Femme	17.9	12e
5984Ho6022002	Togo	Homme	17.7	24e
8778Fo8052006	Togo	Femme	17.59	36e
3869Ho9062007	Togo	Homme	17.51	46e
7341F25122001	Togo	Femme	17.44	60e

e) Construire un nouveau dataframe

Un nouveau DataFrame nommé “ **Admis_candis** ” a été créé afin d’isoler uniquement les candidats ayant réussi le concours. Cette opération permet de faciliter l’analyse des profils admis, en regroupant toutes les informations pertinentes dans un tableau dédié. Ce filtrage est essentiel pour établir des statistiques précises sur les performances des candidats retenus.

IV. Partie 2

1. Vérification de l’assertion ” les pays de l’Afrique de l’Ouest ont des meilleures performances au Bac que les pays de l’Afrique Centrale.”

Une étude menée en 2024 avait affirmé que les candidats d’Afrique de l’Ouest obtenaient de meilleurs résultats au baccalauréat que ceux d’Afrique Centrale. À partir de nos propres données, nous avons souhaité vérifier cette assertion, afin de voir si cette tendance se confirme ou non dans le cadre de notre échantillon.

Pour ce faire, nous avons adopté une approche en plusieurs étapes. Tout d’abord, nous avons enrichi notre base de données en créant deux nouvelles colonnes : **"Mention au Bac"**, qui classe les candidats selon leur performance (Passable, Assez bien, Bien, Très bien), et **"Région"**, qui regroupe les pays en fonction de leur zone géographique, notamment l’Afrique de l’Ouest et l’Afrique Centrale. Par exemple, le Congo est rattaché à l’Afrique Centrale, tandis que le Mali relève de l’Afrique de l’Ouest.

Ces transformations ont permis de préparer les données pour la réalisation d’une **Analyse Factorielle des Correspondances (AFC)**, dont l’objectif était de visualiser les liens entre les mentions obtenues et les régions d’origine des candidats. Enfin, un **test du χ^2 d’indépendance** appuyé par le **V de Cramer** a été appliqué afin de mesurer la significativité statistique des écarts observés.

• Création de la colonne "Mention au Bac"

Dans le cadre de cette étude, une nouvelle colonne intitulée **"Mention Bac"** a été créée afin de classer les candidats selon leur performance au baccalauréat. Cette variable qualitative regroupe les moyennes en différentes catégories telles que *Passable*, *Assez bien*, *Bien* ou *Très bien*. Elle permet de transformer une variable numérique en modalités plus lisibles, facilitant ainsi les comparaisons entre groupes. La création de cette colonne était également indispensable pour réaliser une **Analyse Factorielle des Correspondances (AFC)**, celle-ci ne pouvant être effectuée qu’à partir de variables qualitatives.

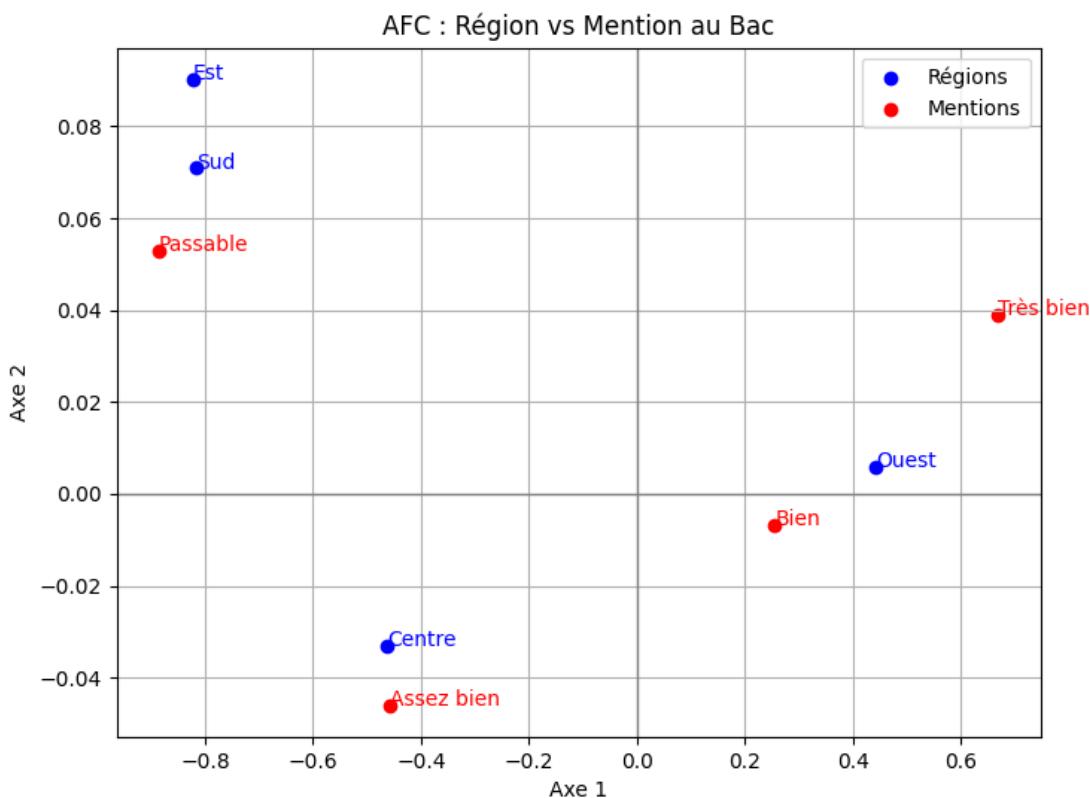
- Analyse Factorielle des Correspondances (AFC)

Afin de réaliser l'**Analyse Factorielle des Correspondances (AFC)**, il est d'abord nécessaire de construire un **tableau de contingence** croisant la variable "**Région**" avec la variable "**Mention au Bac**". Ce tableau permet de comptabiliser le nombre de candidats dans chaque catégorie de mention, selon leur région d'origine.

Une fois ce tableau établi, l'AFC peut être effectuée. Elle aboutit à un **graphique de correspondance** qui permet de visualiser les proximités ou oppositions entre les régions et les mentions obtenues.

Tableau 5: Région x Mention au Bac

Région	Passable	Assez bien	Bien	Très bien
Centre	1870	2161	2910	331
Est	398	344	308	0
Ouest	383	1374	7359	2473
Sud	406	364	319	0



Graphique 4: AFC Mention au bac vs Région

Interprétation

Le graphique d'Analyse Factorielle des Correspondances (AFC) entre la région et la mention au baccalauréat permet de visualiser des regroupements et des oppositions entre certaines catégories, sans toutefois apporter de confirmation statistique.

On observe que la région Ouest se positionne proche des mentions « Bien » et « Très bien », ce qui suggère une association possible entre cette région et les mentions élevées. À l'inverse, les régions Est et Sud apparaissent proches de la mention « Passable », tandis que la région Centre se rapproche de la mention « Assez bien ». Ces positionnements traduisent une tendance visuelle où les meilleures mentions semblent davantage associées à la région Ouest, tandis que les autres régions sont liées à des mentions plus modestes.

Cependant, il faut bien garder à l'esprit que l'AFC ne confirme pas formellement ces liens : elle se limite à offrir une représentation graphique et exploratoire des données, souvent illustrée par un cercle de corrélation. Pour vraiment valider ces observations, il est nécessaire de recourir à des tests statistiques complémentaires, comme le test d'indépendance du khi-deux.

- Test de khi-deux

Ici, nous réalisons le test du khi-deux d'indépendance afin de vérifier si le lien entre la région et la mention au baccalauréat, mis en évidence de manière exploratoire par l'AFC, est statistiquement significatif.

Pour ce faire, nous avons restreint l'analyse aux seules régions Ouest et Centre, afin de vérifier si la supposition de supériorité, visuellement suggérée par l'AFC en faveur de la région Ouest, se confirme statistiquement.

Hypothèses

H_0 : La répartition des mentions au bac est la même pour les candidats de l'Afrique de l'Ouest et de l'Afrique Centrale, autrement dit, la performance au bac ne dépend pas de la région.

H_1 : La répartition des mentions au bac est différente entre les deux régions, autrement dit, la performance au bac dépend de la région d'origine.

Résultat du test du khi-deux

La statistique de test est de 3938,69 avec une p-value très proche de zéro ($p < 0,001$). Cela permet de rejeter l'hypothèse nulle d'indépendance entre la région d'origine (Ouest ou Centre) et la mention au baccalauréat. Autrement dit, les données confirment qu'il existe une association significative entre la région et les performances au bac.

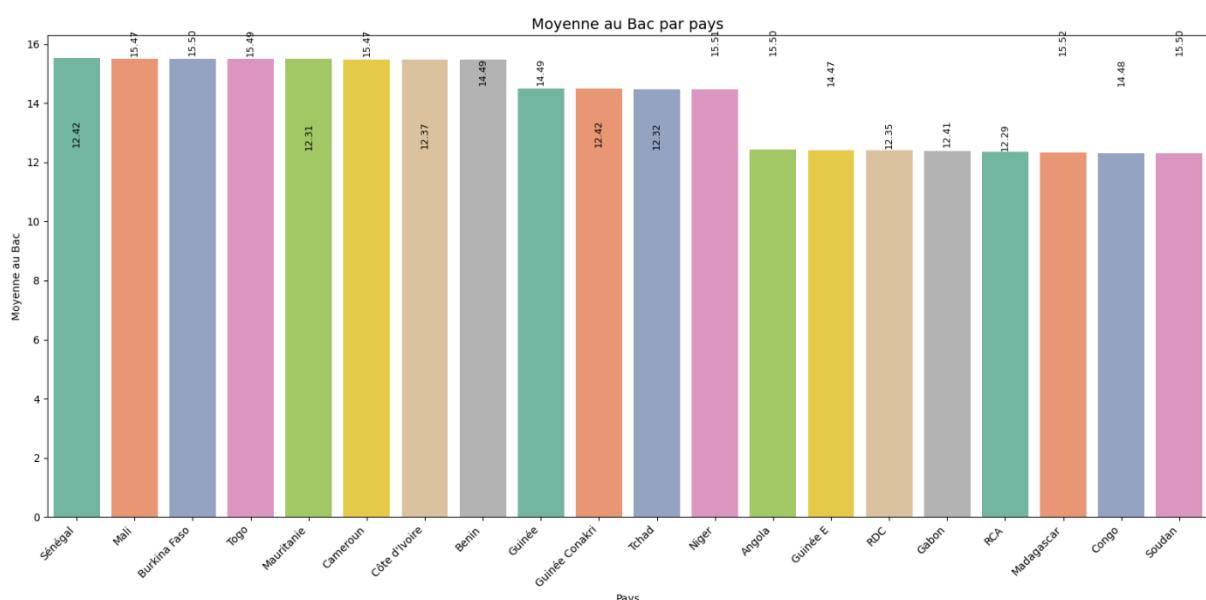
- V de Cramer

Pour mesurer la **force de cette association**, le **V de Cramer** a été calculé et donne une valeur de **0.4570**, indiquant une **association forte** entre la région (Ouest vs Centre) et la mention au Bac.

En conclusion, **ces résultats suggèrent que les candidats originaires de l'Afrique de l'Ouest obtiennent statistiquement de meilleures mentions au Baccalauréat que ceux de l'Afrique Centrale**. Cette conclusion est appuyée à la fois par la **significativité du test du khi-deux** et par la **force du lien mesurée par le V de Cramer**.

2. La performance des candidats au Bac est-elle liée au pays ?

Visualisation des moyennes au Bac par pays



Graphique 5: Répartition de la moyenne au Bac par pays

Interprétation

Ce graphique montre les moyennes au baccalauréat par pays, avec des performances relativement homogènes entre 12 et 16 points. Le Sénégal se distingue avec la moyenne la plus élevée (15,42), suivi du Mali et du Burkina Faso (tous deux à 15,3), tandis que l'Angola affiche la moyenne la plus faible (12,3). La plupart des pays africains représentés maintiennent des moyennes satisfaisantes entre 12 et 15 points, suggérant un niveau d'éducation secondaire globalement correct dans ces systèmes éducatifs.

Pour analyser l'impact du pays sur la performance au Bac, deux approches ont été mobilisées :

- Une première, qualitative, s'est basée sur la variable "*Mention au Bac*". Un test du khi-deux a permis d'évaluer si la distribution des mentions varie significativement selon le pays.
- Une seconde, quantitative, a exploité la variable "*Moyenne au Bac*". Une analyse de variance (ANOVA) a été appliquée afin de comparer les moyennes obtenues entre les pays.

➤ Approche qualitative (Khi-deux plus V de Cramer)

Pour répondre à cette question, nous allons réaliser un **test du Khi-deux d'indépendance** afin de vérifier si la performance des candidats au baccalauréat est liée à leur **pays d'origine**. Étant donné que la variable « Mention au Bac » regroupe les moyennes obtenues en catégories qualitatives, elle constitue un support approprié pour ce type d'analyse. Ce test permettra ainsi de déterminer s'il existe une **association statistiquement significative** entre le pays et la performance au Bac.

Afin d'aller au-delà de la simple existence d'un lien, nous compléterons cette démarche par le **calcul du V de Cramer**, un indicateur de force d'association. Ce dernier nous permettra de mesurer plus précisément **l'intensité** de la relation entre les deux variables, et ainsi d'affiner notre interprétation.

Tableau 6: Mention au Bac vs Pays

Mention au Bac	Assez bien	Bien	Passable	Très bien
Pays				
Angola	367	316	355	0
Benin	0	720	0	331
Burkina Faso	0	719	0	370
Cameroun	0	673	0	330
Congo	342	295	390	0
Côte d'Ivoire	0	765	0	344
Gabon	379	337	406	0
Guinée	336	668	0	2
Guinée Conakri	337	674	0	3
Guinée E	357	344	383	0
Madagascar	364	319	406	0
Mali	0	714	0	343
Mauritanie	0	714	0	377
Niger	344	655	0	1
RCA	334	301	364	0
RDC	384	306	355	0
Soudan	344	308	398	0
Sénégal	0	673	0	358
Tchad	355	682	0	1
Togo	0	713	0	344

Hypothèses

H_0 : La performance au Bac (mention) est indépendante du pays d'origine des candidats.

H_1 : La performance au Bac (mention) dépend du pays d'origine des candidats

Résultat du test du khi-deux et interprétation

La statistique du khi-deux est de 12 869,44 avec une p-value très proche de zéro ($p < 0,001$).

Nous rejetons donc l'hypothèse nulle d'indépendance, ce qui signifie qu'il existe un lien statistiquement significatif entre le pays d'origine et la mention obtenue au bac.

Autrement dit, la performance au bac est bien associée au pays des candidats. Le V de Cramer calculé à 0,4520, indique que cette association est modérée à forte. Cela signifie que le pays d'origine exerce une influence notable sur la performance des candidats au Bac, telle que mesurée par la mention obtenue.

➤ Approche quantitative (ANOVA)

Condition d'application

❖ Normalité des données

La vérification de la normalité des données a été effectuée à l'aide du test de Jarque-Bera, appliqué à la variable "Moyenne au Bac". Les hypothèses de ce test sont :

- H_0 : les données suivent une loi normale ;
- H_1 : les données ne suivent pas une loi normale.

La p-value obtenue est inférieure à 0.05, ce qui nous conduit à rejeter l'hypothèse nulle.

Ainsi, les données ne suivent pas une loi normale, ce qui remet en question l'applicabilité de l'ANOVA classique.

❖ Homoscédasticité des données

Après application du test de Jarque-Bera, les données ne suivent pas une loi normale. Cette violation de la condition de normalité rend l'utilisation de l'ANOVA classique inappropriée.

Par souci de rigueur, nous faisons le test de Levene pour vérifier l'homogénéité des variances entre les groupes.

H_0 : les variances sont égales entre les groupes (homoscédasticité) ;

H_1 : les variances sont différentes entre les groupes (hétéroscédisticité).

Le test de Levene a été réalisé pour vérifier l'homogénéité des variances entre les groupes. Les résultats montrent une p-value inférieure à 0.05, indiquant que la condition d'homoscédasticité n'est pas respectée. Ainsi, les variances des moyennes diffèrent significativement selon les groupes étudiés.

Étant donné que la condition de normalité n'est pas respectée et que l'homogénéité des variances n'est pas vérifiée, l'ANOVA classique ne peut pas être appliquée. Nous allons donc utiliser le test non paramétrique de **Kruskal-Wallis** pour comparer les performances au Bac entre les différents pays.

Test de Kruskal-Wallis

Hypothèses

Ho : la distribution des moyennes au Bac est identique quel que soit le pays d'origine des candidats.

H₁ : au moins un pays présente une distribution différente des autres.

Compte tenu du non-respect des conditions de normalité et d'homogénéité des variances, nous avons appliqué le test non paramétrique de **Kruskal-Wallis** pour évaluer si la moyenne au Bac varie selon le pays d'origine des candidats.

Le test a donné une **statistique de 12034.03** avec **une p-value de 0.0000**, indiquant une différence hautement significative entre les pays.

Conclusion : la moyenne au Bac des candidats varie significativement en fonction de leur pays d'origine.

Analyse post-hoc (test de Dunn)

Résultats du test de Dunn post-hoc (correction de Bonferroni)				
	Angola	Benin	Burkina Faso	Cameroun
Angola	1.000000e+00	1.405068e-307	0.000000e+00	3.713410e-304
Benin	1.405068e-307	1.000000e+00	1.000000e+00	1.000000e+00
Burkina Faso	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
Cameroun	3.713410e-304	1.000000e+00	1.000000e+00	1.000000e+00
Congo	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
Côte d'Ivoire	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
Gabon	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
Guinée	1.172985e-100	2.029122e-53	4.404253e-58	9.632570e-54
Guinée Conakri	2.857528e-100	3.198432e-54	6.148193e-59	1.549466e-54
Guinée E	1.000000e+00	0.000000e+00	0.000000e+00	2.698066e-308
Madagascar	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
Mali	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
Mauritanie	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
Niger	9.507378e-96	1.008914e-56	1.522796e-61	5.101722e-57
RCA	1.000000e+00	2.424399e-305	0.000000e+00	4.667631e-302
RDC	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
Soudan	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
Sénégal	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
Tchad	3.017266e-99	1.561218e-56	2.110045e-61	8.040804e-57
Togo	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
	Congo	Côte d'Ivoire	Gabon	Guinée
Angola	1.000000e+00	0.000000e+00	1.000000e+00	1.172985e-100
...				
Soudan	1.000000e+00	0.000000e+00	3.040331e-107	0.000000e+00
Sénégal	0.000000e+00	1.000000e+00	3.387199e-62	1.000000e+00
Tchad	3.040331e-107	3.387199e-62	1.000000e+00	2.848400e-61
Togo	0.000000e+00	1.000000e+00	2.848400e-61	1.000000e+00

Suite au test de Kruskal-Wallis, une analyse post-hoc a été réalisée à l'aide du **test de Dunn**, avec une correction de Bonferroni pour comparer les moyennes au Bac entre tous les pays deux à deux. Les résultats ont mis en évidence **plusieurs différences significatives** entre pays, notamment entre l'**Angola**, le **Congo**, le **Gabon**, le **Niger**, la **Guinée**, le **Tchad**, et plusieurs autres. Ces différences confirment que la **moyenne au Bac varie effectivement selon le pays d'origine des candidats**, donc le pays d'origine a un impact sur la moyenne au bac.

3. Lien entre la performance au Bac et la région

Pour évaluer si la performance des candidats au baccalauréat varie selon leur région d'origine, nous avons appliqué **un test du Khi-deux d'indépendance**, adapté à l'analyse de deux variables qualitatives : la région et la mention au Bac. Afin de compléter l'interprétation, nous avons également calculé **le V de Cramer**, qui permet de mesurer l'intensité de l'association.

Tableau 7: Mention au Bac vs Région

Mention au Bac	Assez bien	Bien	Passable	Très bien
Région				
Centre	2161	2910	1870	331
Est	344	308	398	0
Ouest	1374	7359	383	2473
Sud	364	319	406	0

Hypothèses formulées

- H_0 : Il n'existe aucune association entre la région d'origine et la performance au Bac.
- H_1 : Il existe une association significative entre la région d'origine et la performance au Bac.

Résultats du test

- Statistique du Khi-deux : **5280.86**
- p-value : < **0.0001**
- V de Cramer : **0.29**

Interprétation

La p-value étant largement **inférieure à 0,05**, nous rejetons l'hypothèse nulle. Il existe donc une association statistiquement significative entre la région et la mention obtenue au baccalauréat. Le **V de Cramer**, avec une valeur de **0,29**, indique une

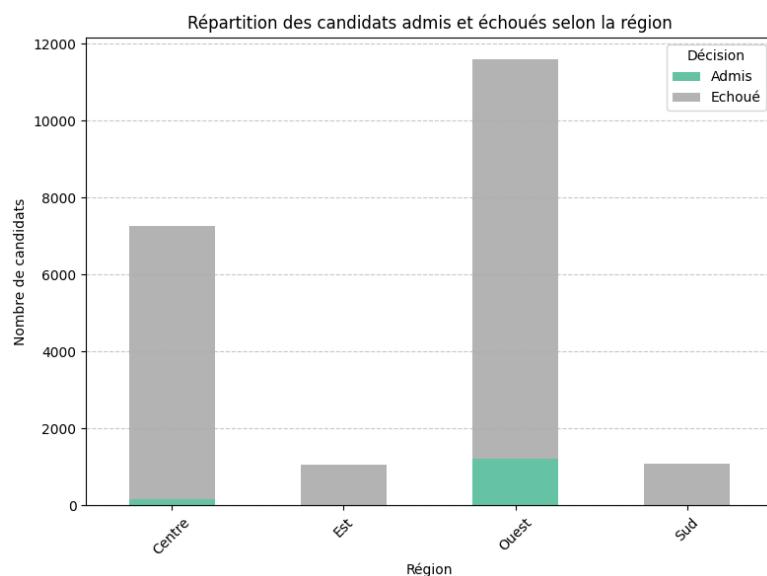
liaison modérée, suggérant que la région peut avoir une influence sur les performances, sans en être le seul facteur explicatif.

4. Lien entre la région et l'admission au concours

Pour savoir si l'admission au concours EAMAC varie selon la région d'origine des candidats, nous avons utilisé le **test du Khi-deux d'indépendance**, qui permet d'évaluer l'existence d'un lien entre deux variables qualitatives : **la région** et **la décision d'admission** (Admis / Échoué). L'analyse est complétée par le **V de Cramer**, afin de mesurer la force de l'association.

Tableau 8: Région vs Admission au concours

Décision	Admis	Echoué
Région		
Centre	155	7117
Est	1	1049
Ouest	1223	10366
Sud	2	1087



Graphique 6: Répartition des candidats admis et échoués selon leur région d'origine

Interprétation du graphique

La région Ouest se distingue par un effectif très élevé de candidats et une proportion significative d'admis, ce qui suggère une meilleure performance globale. En revanche, bien que la région Centre présente également un grand nombre de candidats, la proportion d'admis y est nettement plus faible, laissant entrevoir des difficultés particulières. Les régions Est et Sud affichent un effectif réduit et très peu d'admis, ce qui pourrait s'expliquer par un faible nombre de participants ou des résultats plus modestes. L'ensemble du graphique laisse donc présager un lien possible entre la région d'origine et les chances de réussite au concours, avec un net avantage observé pour la région Ouest.

Hypothèses du test :

- H_0 : Il n'existe aucun lien entre la région d'origine et l'admission au concours.
- H_1 : Il existe une association significative entre la région d'origine et l'admission au concours.

Résultats du test et interprétation

Le test du Khi-deux réalisé entre la région d'origine des candidats et leur admission au concours donne une statistique de 676,39 avec une p-value inférieure à 0,0001. Ce résultat nous conduit à rejeter l'hypothèse nulle, ce qui signifie qu'il existe une association statistiquement significative entre la région et la décision d'admission. Pour apprécier l'intensité de ce lien, nous avons calculé le V de Cramer, qui s'élève à 0,1795. Cette valeur indique une liaison de faible à modérée : la région semble donc jouer un rôle dans les résultats d'admission, sans en être l'unique facteur déterminant.

5. Lien entre l'admission au concours et la moyenne au Bac

Pour analyser la relation entre la moyenne obtenue au baccalauréat et l'admission au concours, nous avons eu recours à une régression logistique binaire. Cette méthode est particulièrement adaptée lorsque la variable dépendante est dichotomique « Décision » prenant les modalités « Admis » ou « Échoué » et que l'on souhaite étudier l'effet d'une variable explicative quantitative, en l'occurrence la moyenne au Bac.

Étant donné que nos données ne suivent pas une loi normale, et que la variable explicative est quantitative tandis que la variable à expliquer est qualitative binaire, les tests paramétriques classiques tels que l'ANOVA ne sont pas appropriés.

Dans ce contexte, la régression logistique apparaît comme le test statistique le plus pertinent pour répondre à notre problématique.

Avant de procéder à l'analyse, la variable « Décision » a été recodée en variable binaire, et les modalités ont été préalablement normalisées afin de garantir la qualité et la cohérence des données.

Afin de cadrer notre analyse, nous avons formulé les hypothèses suivantes :

- **H₀** : la moyenne au Bac n'a aucun effet sur la décision d'admission au concours.
- **H₁** : la moyenne au Bac a un effet significatif sur la décision d'admission au concours.

Iterations 8							
Logit Regression Results							
Dep. Variable:	Admission_binaire	No. Observations:	21000 <th>Model:</th> <th>Logit</th> <th>Df Residuals:</th>	Model:	Logit	Df Residuals:	
Method:	MLE	Df Model:	1	Date:	Fri, 11 Jul 2025	Pseudo R-squ.:	0.07093
Time:	23:28:11	Log-Likelihood:	-4732.0	converged:	True	LL-Null:	-5093.2
Covariance Type:	nonrobust	LLR p-value:	3.693e-159				
	coef	std err	z	P> z	[0.025	0.975]	
const	-10.0933	0.336	-30.018	0.000	-10.752	-9.434	
...							
Odds ratios :							
const	0.000041						
Moyenne au Bac	1.659121						

Interprétation

Nous avons effectué une régression logistique pour évaluer l'impact de la moyenne au baccalauréat sur la probabilité d'admission au concours. La variable dépendante est binaire (Décision : admis = 1, échoué = 0) et la variable explicative est la moyenne au Bac (quantitative).

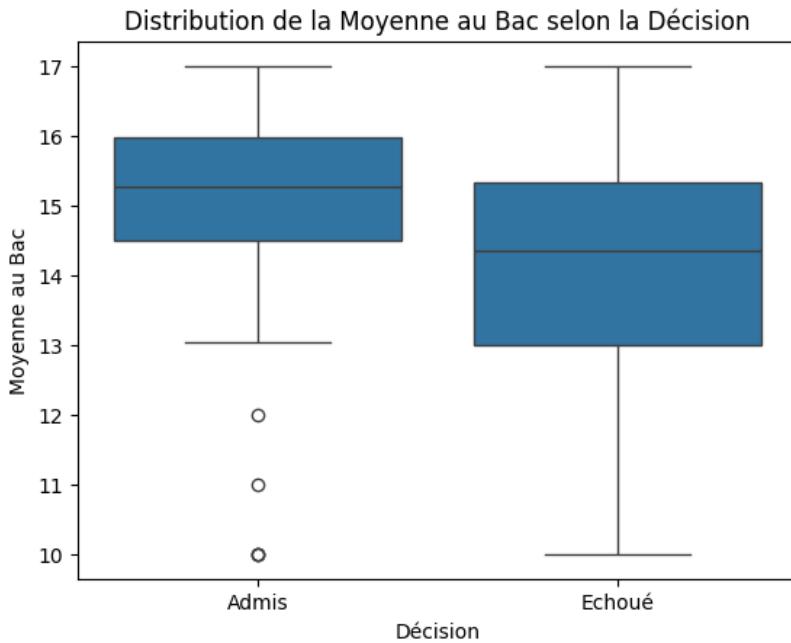
Le modèle converge correctement avec un effet significatif de la moyenne au Bac (p-value < 0,001). Le coefficient positif indique que plus la moyenne au Bac est élevée, plus la probabilité d'être admis augmente.

L'odds ratio estimé est d'environ 1,66, ce qui signifie que chaque point supplémentaire à la moyenne au Bac multiplie par 1,66 la chance d'admission au concours, toutes choses égales par ailleurs.

Ce résultat confirme donc que la moyenne au Bac influence significativement l'admission au concours.

Visualisation de la relation entre la moyenne au Bac et l'admission

Le boxplot ci-dessous compare la moyenne au Bac entre les candidats **admis** et **échoués**.



Graphique 7: Moyenne au Bac vs Admission au concours

Interprétation

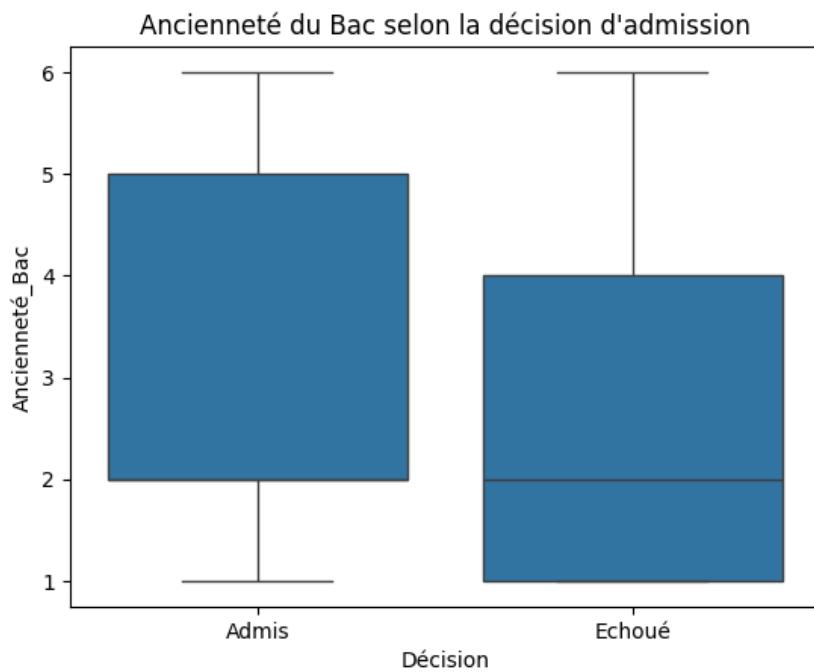
Ce graphique montre que les admis présentent, en général, des moyennes plus élevées, avec une médiane supérieure. Cette représentation visuelle confirme la tendance observée dans la régression logistique : **une meilleure moyenne au Bac est associée à une probabilité plus élevée d'admission au concours.**

6. Lien entre l'ancienneté du bac et l'admission au concours

Dans l'opinion de certains, l'admission au concours serait influencée par la maturité du candidat, c'est-à-dire par l'ancienneté de son baccalauréat. Selon cette hypothèse, les candidats ayant obtenu leur Bac depuis plusieurs années auraient davantage de chances d'être admis. Afin de vérifier cette affirmation, nous avons créé une variable nommée **Ancienneté_Bac**, correspondant au nombre d'années écoulées entre l'obtention du Bac et l'année 2025. Dans le but de comprendre si l'ancienneté du

baccalauréat influence la réussite au concours, nous avons choisi d'utiliser une régression logistique simple. Cette méthode permet de modéliser la probabilité d'admission (variable binaire) en fonction de l'ancienneté, sans nécessiter que les données suivent une loi normale. Ainsi, elle est parfaitement adaptée à notre problématique et à la nature de nos données. Cette approche statistique nous permettra de quantifier l'impact éventuel de l'ancienneté sur la réussite au concours.

Visualisation



Graphique 8: Ancienneté du bac vs admission au concours

Interprétation

Ce graphique révèle une observation surprenante : les candidats admis semblent avoir un Bac plus "ancien" (médiane ~3,5 ans) que ceux qui ont échoué (médiane ~2 ans). À première vue, cela va à l'encontre de l'idée intuitive qu'un diplôme récent serait un avantage. Toutefois, cette différence visuelle pourrait être trompeuse et il faudrait donc vérifier avec des tests statistiques (régression logistique) si cette tendance est réellement significative ou juste due au hasard.

Hypothèses du test

- **H₀** : L'ancienneté du baccalauréat n'a aucun effet sur la probabilité d'admission au concours.
- **H₁** : L'ancienneté du baccalauréat influence la probabilité d'admission au concours.

Logit Regression Results						
Dep. Variable:	Admission_binaire	No. Observations:	21000			
Model:	Logit	Df Residuals:	20998			
Method:	MLE	Df Model:	1			
Date:	Sat, 12 Jul 2025	Pseudo R-squ.:	7.676e-07			
Time:	07:52:04	Log-Likelihood:	-5093.2			
converged:	True	LL-Null:	-5093.2			
Covariance Type:	nonrobust	LLR p-value:	0.9295			
	coef	std err	z	P> z	[0.025	0.975]
const	-2.6494	0.056	-47.733	0.000	-2.758	-2.541
Ancienneté_Bac	-0.0014	0.016	-0.088	0.930	-0.033	0.030
const	0.070691					
Ancienneté_Bac	0.998583					

Interprétation du modèle de régression logistique

Nous avons réalisé une régression logistique simple pour étudier l'effet de l'ancienneté du candidat au baccalauréat (Ancienneté_Bac) sur la probabilité d'admission au concours.

Les résultats montrent que le coefficient estimé pour l'ancienneté est de **-0,0014** ($p = 0,93$), ce qui indique que l'ancienneté n'a pas d'impact statistiquement significatif sur la probabilité d'admission. L'odds ratio associé est de **0,999**, très proche de 1, signifiant qu'une année supplémentaire d'ancienneté modifie de manière négligeable la probabilité d'admission.

Le modèle, basé uniquement sur cette variable explicative, ne présente pas de pouvoir explicatif pertinent (Pseudo R^2 quasi nul), ce qui signifie que l'ancienneté prise isolément ne permet pas de prédire l'admission au concours.

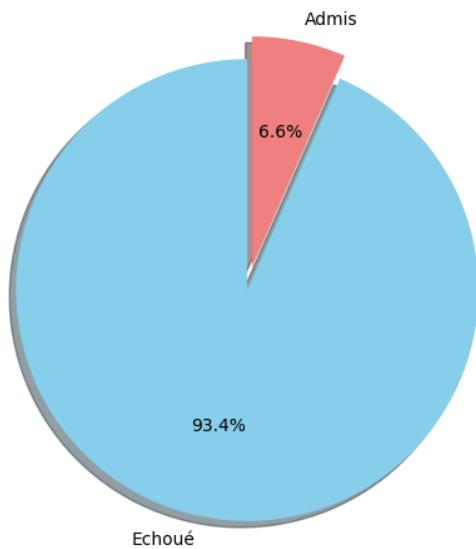
En conclusion, dans notre échantillon, l'ancienneté du candidat au baccalauréat ne semble pas être un facteur déterminant pour l'admission au concours.

7. Modélisation par Machine Learning

Avant de modéliser un modèle de machine learning pour prédire l'admission au concours, il est important de visualiser la part des admis et des échoués .

➤ Visualisation des admis et des échoués au concours

Répartition des admis et des échoués au concours



Graphique 9: Répartition des candidats admis et échoués au concours

Interprétation

Ce graphique met en évidence que seuls 6,6 % des candidats au concours ont été déclarés admis, contre une large majorité de 93,4 % d'échecs. Cela révèle un taux de sélection extrêmement rigoureux, témoignant d'un niveau d'exigence très élevé. Le concours apparaît ainsi comme hautement compétitif, réservé à une élite académique.

Modèle de classification (Régression logistique)

Afin de prédire l'admission des candidats au concours de l'EAMAC, nous avons conçu un modèle de machine learning de **régression logistique**, fondé sur un ensemble de variables **académiques** (notes individuelles au concours, moyenne au bac, ...) et **sociodémographiques** (Age, Sexe, Région, etc.). Le prétraitement des données a reposé sur une **standardisation des variables numériques** et un **encodage one-hot des variables qualitatives**. Face à l'important **déséquilibre des classes** (seulement 6,6 % de candidats admis), nous avons exploré l'usage de **SMOTE**, une technique de suréchantillonnage synthétique, pour améliorer la détection de la classe minoritaire. Toutefois, pour préserver la validité du **test de**

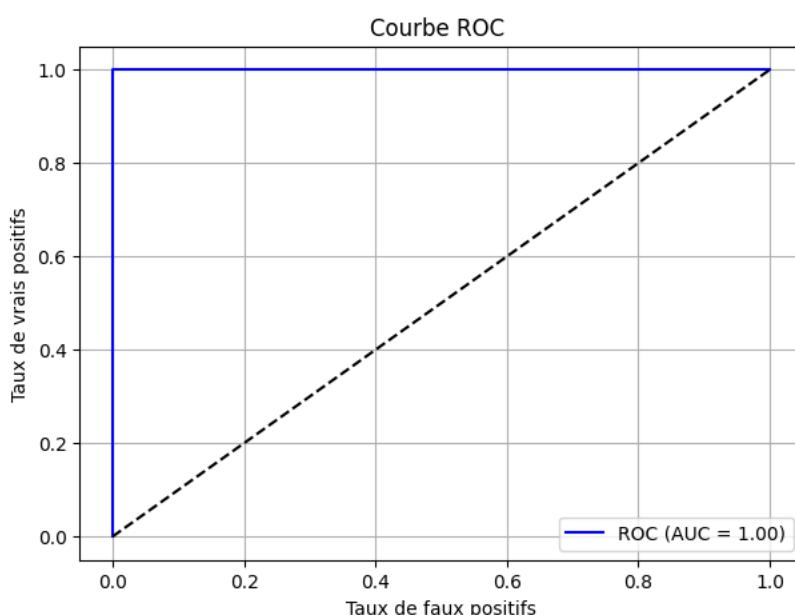
Hosmer-Lemeshow, crucial dans l'évaluation de la **calibration du modèle**, la version finale a été entraînée **sans SMOTE**, sur les données brutes.

Tout au long du processus, nous avons porté une attention particulière à la **robustesse** du modèle. Étonnés par les **résultats remarquablement élevés** (accuracy > 99 %, AUC ≈ 1, Brier score ≈ 0.003), nous avons vérifié la présence d'un éventuel **surapprentissage**. Les **performances stables** sur les jeux d'entraînement et de test ont confirmé que le modèle **généralise bien** sur de nouvelles données. Le **test de Hosmer-Lemeshow** ($p = 0.94$) a validé une **calibration optimale**, confirmant que le modèle reflète fidèlement les probabilités réelles d'admission.

8. Vérification de la robustesse du modèle

Pour évaluer la robustesse de notre modèle, nous avons analysé:

- La courbe ROC et le taux de bon classement



Graphique 10: Courbe Roc

Interprétation

La courbe ROC ci-dessus témoigne de l'excellente performance du modèle de régression logistique. Avec une surface sous la courbe (AUC) égale à 1.00, le modèle montre une capacité parfaite à distinguer les candidats admis de ceux échoués. Cette courbe colle au coin supérieur gauche du graphique, ce qui indique un très bon taux de vrais positifs et un faible taux de faux positifs. En comparaison, la diagonale noire représente un modèle aléatoire ($AUC = 0.5$), bien moins performant. Ainsi, la forme et la position de la courbe ROC confirment que notre modèle est à la fois discriminant, fiable, et bien calibré sur les données disponibles.

- Test de HosmerLemeshow

Afin d'évaluer la calibration de notre modèle de régression logistique, nous avons appliqué le test de Hosmer-Lemeshow, couramment utilisé dans la littérature biométrique et statistique. Ce test permet de vérifier si les probabilités prédites par le modèle sont cohérentes avec les observations réelles.

Hypothèses

H_0 : Le modèle est bien calibré, c'est-à-dire que les probabilités prédites (admis ou échoué) reflètent correctement les probabilités observées.

H_1 : Le modèle est mal calibré, les probabilités prédites ne reflètent pas la réalité.

Résultats du test

Le test de Hosmer-Lemeshow, appliqué à notre modèle de régression logistique, a donné une **statistique de 2.8661** avec une **p-value de 0.9425**. Étant donné que cette p-value est **largement supérieure au seuil conventionnel de 0.05**, nous **n'avons pas de raison de rejeter l'hypothèse nulle**. Cela signifie que les probabilités prédites par le modèle sont cohérentes avec les données observées : autrement dit, **le modèle est bien calibré**.

Cette bonne calibration confirme que notre modèle ne se contente pas de bien classifier, mais qu'il estime également **de manière fiable les chances d'admission** pour chaque candidat.

V. Conclusion

Cette étude approfondie des données d'admission au concours EAMAC 2025 a permis de mettre en lumière plusieurs éléments clés concernant les dynamiques de sélection et les facteurs influençant la réussite des candidats.

L'analyse des 21 000 candidats révèle d'abord une forte sélectivité du concours, avec seulement 6,6% d'admis, confirmant le caractère d'élite de cette formation. Cette rigueur dans la sélection, bien que nécessaire pour maintenir la qualité, soulève des questions importantes sur l'équité d'accès à cette formation stratégique pour le continent africain.

Sur le plan géographique, les résultats confirment partiellement les hypothèses de départ. L'assertion selon laquelle les candidats d'Afrique de l'Ouest obtiendraient de meilleures performances au baccalauréat que ceux d'Afrique Centrale s'est révélée statistiquement fondée, avec un V de Cramer de 0,457 témoignant d'une association forte. Cette tendance se traduit effectivement par une meilleure représentation des pays ouest-africains parmi les admis, le Mali (18,3%), le Sénégal (17,7%) et la Côte d'Ivoire (17,1%) dominant le classement des taux de réussite.

Cependant, l'analyse révèle aussi des disparités préoccupantes. Certains pays comme le Gabon, le Congo, Madagascar ou la RCA affichent des taux de réussite quasi nuls, soulevant des interrogations sur l'équité du processus et l'égalité des chances selon l'origine géographique. Ces écarts ne peuvent s'expliquer uniquement par les différences de niveau, mais suggèrent des inégalités plus profondes dans l'accès à l'information, la préparation ou même les conditions socio-économiques.

L'étude a également démontré l'impact significatif de la moyenne au baccalauréat sur l'admission au concours, avec un odds ratio de 1,66, confirmant que chaque point supplémentaire au bac multiplie considérablement les chances de réussite. En revanche, contrairement aux idées reçues, l'ancienneté du baccalauréat ne constitue pas un facteur déterminant, remettant en question l'hypothèse selon laquelle la maturité académique favoriserait l'admission au concours.

Le modèle de machine learning de régression logistique développé pour prédire l'admission au concours a obtenu une AUC parfaite de 1,00 et une excellente calibration selon le test de Hosmer-Lemeshow. Ces performances, bien que statistiquement impressionnantes, soulèvent des questions sur la rigidité du processus de sélection et sa capacité à refléter la diversité des talents car parfois les

meilleurs candidats sont souvent ceux qui sortent du moule standard et apportent une valeur ajoutée inattendue.

Limites de l'étude

Malgré la richesse des données, plusieurs points méritent d'être soulignés pour bien interpréter nos résultats.

- **Représentativité inégale :** Tous les pays africains ne sont pas représentés de façon équilibrée. Certains ont très peu de candidats, ce qui rend les comparaisons délicates. Cette disparité révèle probablement des inégalités d'accès à l'information ou des obstacles logistiques variables selon les régions.
- **Vision partielle :** Notre analyse se limite aux données chiffrées disponibles. On ne connaît ni le contexte socio-économique des candidats, ni leur accès aux formations préparatoires, ni la qualité de leurs établissements d'origine. Ces facteurs pourraient pourtant expliquer une bonne part des écarts observés.
- **Regard sur une seule année :** Cette étude porte uniquement sur 2025 et ne permet pas de dégager des tendances. Les performances d'un pays peuvent fluctuer selon les réformes éducatives, le contexte politique ou d'autres facteurs conjoncturels.
- **Format standardisé :** Le concours, avec ses quatre épreuves uniformes, privilégie peut-être certains profils académiques. D'autres formes d'intelligence ou de compétences, pourtant essentielles en météorologie et aviation civile, pourraient être négligées.

Recommandations

Face à ces constats, en tant que data scientist, voici les recommandations formulées à l'intention de l'EAMAC afin d'optimiser les chances d'admission des candidats au concours.

- **Réduire les inégalités géographiques** : Mener une enquête dans les pays sous-représentés pour comprendre les obstacles (manque d'info, difficultés d'accès, formations inadéquates). Puis développer des solutions ciblées : campagnes d'information, sessions de préparation délocalisées, partenariats avec les écoles locales.
- **Envisager des quotas régionaux** : Garantir une représentation minimale de chaque région sans baisser le niveau d'exigence. Une mesure délicate à mettre en place mais qui assurerait plus d'équité.
- **Élargir les critères d'évaluation** : Garder les épreuves académiques mais y ajouter d'autres dimensions : compétences pratiques, etc. Cela révélerait des profils complémentaires.
- **Créer un observatoire des admissions** : Suivre les tendances sur plusieurs années pour détecter rapidement les déséquilibres et ajuster les stratégies.
- **Améliorer l'accompagnement** : Développer une plateforme en ligne avec ressources pédagogiques, annales et conseils. Cela démocratiserait l'accès à une bonne préparation.
- **Développer des partenariats** : Collaborer avec les institutions des pays moins représentés via des échanges pédagogiques et du mentorat.