Nom et Prénom :

CY TECH	**ING1 (GIA-GMA)** EXAMEN DE DATA EXPLORATION 2023-2024
2h	2 feuilles R/V (manuscrites ou non) autorisées Calculatrice autorisée

Ce sujet contient un seul exercice avec 5 parties indépendantes.

« Les rêves ne sont pas ce que vous voyez dans votre sommeil, les rêves sont des choses qui ne vous permettent pas de dormir. » Cristiano Ronaldo

La Coupe du monde de football (FIFA), est le championnat du monde des équipes nationales masculines de football. Décidée le 28 mai 1928 par la Fédération internationale de football association (FIFA) sous l'impulsion de son président Jules Rimet, elle a été ouverte à toutes les équipes des fédérations reconnues par la FIFA, professionnelles y compris, se distinguant en cela du tournoi olympique de football, à l'époque réservé aux amateurs. Elle a lieu pour la première fois en 1930, en Uruguay, et tous les quatre ans depuis (sauf en 1942 et 1946 en raison de la Seconde Guerre mondiale).

Le jeu de données **de l'Annexe** « FIFA» contient les informations sur chaque tournoi de la Coupe du monde de 1930 à 2022. Les variables étudiées sont :

- Année : L'année du tournoi.
- Pays : Le pays hôte du tournoi.
- Vainqueur : L'équipe qui a remporté le tournoi.
- GoalsScored : Le nombre total de buts marqués dans le tournoi.
- QualifiedTeams : Le nombre total d'équipes qualifiées pour le tournoi.
- MatchesPlayed : Le nombre des matchs joués.
- Attendance : Le nombre total de spectateurs qui ont assisté aux matches.

A- Analyse Univariée

1. Quelle est la population étudiée? Quelle est la taille de l'échantillon?

2. Donner le type des variables étudiées en indiquant comment peut-on les représenter graphiquement.

3. On donne les résumés numériques de la variable « **Attendance** » (en milliers de spectateurs).

```
>summary(Attendance)
   Min.   1st Qu.  Median    Mean   3rd Qu.   Max.
  81.98   799.75  1987.74  1898.41  2970.10  3587.54
```

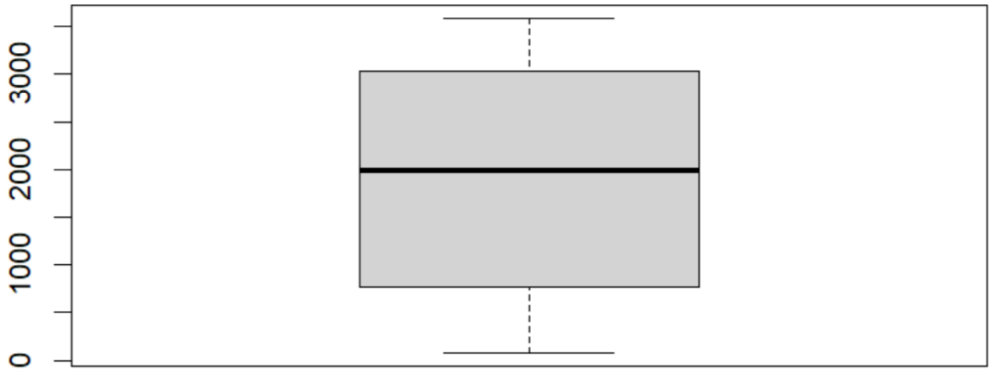
a) Donnez les quartiles et interprétez les.

b) Calculez l'écart interquartiles et l'étendue.

c) Quel est le nombre moyenne des spectateurs par match? Comparez cette valeur avec la médiane.

d) Parmi les indicateurs numériques calculés avant, lesquelles sont des indicateurs de dispersion et lesquels sont des indicateurs de positions.

4. On donne le Boxplot de la variable « Attendance».



a) Sur le graphique de Boxplot, placez Q1, Q2 et Q3.

b) Calculez les moustaches du boxplot. Interprétez ces deux valeurs et notez-les sur le graphique.

c) Y a-t-il des valeurs aberrantes pour la variable Attendance ? Justifiez votre réponse.

B- Croisement Quantitative x Quantitative

Notre objectif est de prédire le nombre des matchs joués Y (MatchesPlayed) en fonction des équipes qualifiées X (QualifiedTeams). Ci-dessous les résultats fournis par le logiciel R.

```
> summary(ModelSimple)
Coefficients:
                  Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)        -5.4555     3.5721     -1.527     0.142
QualifiedTeams      2.2168      0.1524     14.544    4.25e-12 ***
---
Multiple R-squared:  0.9136,       Adjusted R-squared:  0.9093
F-statistic: 211.5 on 1 and 20 DF,            p-value: 4.248e-12
```

1. Précisez la variable explicative et la variable à expliquer.

2. Déterminez la droite de régression de Y en X.

3. Déterminez le coefficient de détermination. Interprétez.

4. Déterminez le coefficient de corrélation. Interprétez.

5. Prédire le nombre des matches joués (MatchesPlayed) si on a 32 équipes qualifiées.

6. On donne les résidus pour chaque observation :

```
> Modelsimple$residuals
      1       2       3       4       5       6       7
-5.363041 -13.013481 -9.796667 -1.363041 -4.013481  4.986519  1.986519
      8       9      10      11      12      13      14
 1.986519  1.986519  7.986519  7.986519  4.252013  4.252013  4.252013
     15      16      17      18      19      20      21      22
 4.252013 -1.482494 -1.482494 -1.482494 -1.482494 -1.482494 -1.482494 -1.482494
```

a) Justifiez avec les calculs pourquoi le résidu de la première observation est e_1= -5.363041.

b) Peut-on détecter les valeurs aberrantes avec ces simples résidus? justifiez votre réponse.

C - Croisement Qualitative x Qualitative

A partir des données, on considère maintenant les deux variables qualitatives :

X : représente le continent hôte du tournoi (Afrique, Amérique, Asie, Europe).

Y : représente le continent du pays qui a gagné le tournoi (Amérique, Europe).

On souhaite alors étudier le lien entre X et Y, on donne le tableau suivant.

Y \X	Aferique	Amerique	Asie	Europe
Amerique	0	7	1	2
Europe	1	1	2	8

1. Complétez les cases vides dans le tableau suivant.

Y \X	Aferique	Amerique	Asie	Europe	Total
Amerique	0	7	1	2	
Europe	1	1	2	8	
Total					

2. Complétez le tableau des fréquences observées.

Y \X	Aferique	Amerique	Asie	Europe	Total
Amerique					
Europe					
Total					

3. Calculez le tableau des profils lignes.

Y \X	Aferique	Amerique	Asie	Europe
Amerique				
Europe				

4. Calculez le tableau des profils colonnes.

Y \X	Aferique	Amerique	Asie	Europe
Amerique				
Europe				

5. Donnez et interprétez les valeurs suivantes.

n_{22} :

$n_{.2}$:

$n_{2.}$:

f_{21} :

$f_{3/2}$:

$f_{2/3}$:

$f_{2/2}$:

6. Donnez le profil moyenne ligne et interprétez ces valeurs.

7. Donnez le profil moyenne colonne et interprétez ces valeurs.

8. Complétez le tableau des effectifs théoriques.

Y \X	Aferique	Amerique	Asie	Europe	Total
Amerique					
Europe					
Total					

9. Sachant que le khi deux vaut ici 9.3286. Etudiez le lien entre les deux variables X et Y.

Voici le tableau de seuils de décision

d.d.l	1	2	3	4	5	6	7	8	9	10
Seuil	3.84	5.99	7.82	9.49	11.075	12.59	14.07	15.51	16.92	18.31

C - Croisement Quantitative x Qualitative

Maintenant, on étudie le lien entre la variable quantitative QualifiedTeams et la variable X qui représente le continent hôte du tournoi et qui a 4 modalités (Afrique, Amérique, Asie, Europe).

Le logiciel R a nous fournit les résultats suivants :

```
> Model2 <- lm(WorldCups$QualifiedTeams ~as.factor(WorldCups$Conti.hote))
> anova(Model2)
Analysis of Variance Table
Response: WorldCups$QualifiedTeams
                                Df  Sum Sq  Mean Sq  F value  Pr(>F)
as.factor(WorldCups$Conti.hote)  3  476.26  158.755   3.8742  0.02675 *
Residuals                       18  737.60   40.978
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

1. Calculez la variance inter-groupe.

2. Calculez la variance intra groupe.

3. Déduire la variance totale.

4. Calculez et interprétez le rapport de corrélation.

D – Analyse en Composantes Principales (ACP)

Nous réalisons une analyse en composantes principales sur les variables suivantes : Attendance, QualifiedTeams, GoalsScored et MatchesPlayed.

1. Quel est le but de réaliser un ACP.
2. Pourquoi faut-il centrer et réduire les variables avant d'appliquer une analyse multivariée?
3. De combien des points a-t-il constitué le nuage du point?
4. Combien de dimensions sont nécessaires pour représenter ces points ?

On donne les résultats de l'ACP fournit par le logiciel R.

```
> ACP$eig
        eigenvalue     percentage of variance     cumulative percentage of variance
comp 1  3.63978175          90.994544                  90.99454
comp 2  0.23384990          5.846247                   96.84079
comp 3        a             2.068867                   98.90966
comp 4   0.04361367            b                      100.00000

> ACP$var$cos2
                 Dim.1      Dim.2          Dim.3       Dim.4
GoalsScored   0.8656987  1.104077e-01   0.02377399  0.0001196404
```

QualifiedTeams	0.9552522	1.688411e-04	0.02630803	0.0182709243
MatchesPlayed	0.9615642	1.070264e-05	0.01339465	0.0250304406
Attendance	0.8572666	1.232627e-01	0.01927802	0.0001926663

> ACPvarcontrib

	Dim.1	Dim.2	Dim.3	Dim.4
GoalsScored	23.78436	47.2130	28.7282	0.2743185
QualifiedTeams	26.24477	0.07220	31.79038	41.8926535
MatchesPlayed	26.41818	0.00457	16.18597	57.3912713
Attendance	23.55269	52.7101	23.29538	0.4417567

5. Utiliser les sorties de R pour répondre aux questions suivantes.

a) Comment est définie la quantité d'information (inertie) contenue dans le nuage de points? A partir de quelle matrice peut-on la calculer et comment la calcule-t-on?

b) Calculez les deux valeurs manquantes a et b dans les sorties R.

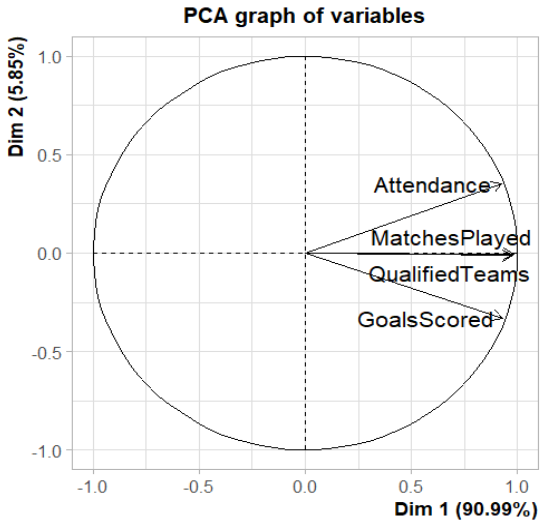
c) Combien d'axes faut-il retenir et pourquoi ?

d) Donnez la contribution moyenne de chaque variable.

e) Donnez la contribution moyenne de chaque individu.

f) Sur quels axes la variable Matchesplayed sera le mieux représenté. Justifiez votre réponse.

6) Interprétez le graphe des variables



Pensez à relire! Bon courage

"Soon, when all is well, you're going to look back on this period of your life and be so glad that you never gave up."

ANNEXE

	Country	Winner	GoalsScored	QualifiedTeams	MatchesPlayed	Attendance
1930	Uruguay	Uruguay	70	13	18	590549
1934	Italy	Italy	70	16	17	363000
1938	France	Italy	84	15	18	376000
1950	Brazil	Uruguay	88	13	22	1045246
1954	Switzerland	Germany	140	16	26	768607
1958	Sweden	Brazil	126	16	35	81981
1962	Chile	Brazil	89	16	32	893172
1966	England	England	89	16	32	1563135
1970	Mexico	Brazil	95	16	32	1603975
1974	Germany	FR	97	16	38	1865753
1978	Argentina	Argentina	102	16	38	154791
1982	Spain	Italy	146	24	52	2109723
1986	Mexico	Argentina	132	24	52	2394031
1990	Italy	FR	115	24	52	2516215
1994	USA	Brazil	141	24	52	3587538
1998	France	France	171	32	64	2785100
2002	Korea/Japan	Brazil	161	32	64	2705197
2006	Germany	Italy	147	32	64	3359439
2010	Africa	Spain	145	32	64	3178856
2014	Brazil	Germany	171	32	64	3386810
2018	Russie	France	169	32	64	3031768
2022	Qatar	Argentine	172	32	64	3404052