Projet Statistiques descriptives avec R – à rendre pour le 16 mai 2024

Vous rendrez deux fichiers par mail à l'adresse `christophe.ambroise@univ-evry.fr`: un fichier `nom-prenom.rmd` avec votre code commenté et un fichier pdf ou word, version compilée du premier fichier.

Exercice 1

1. Charger le jeu de données `lex.csv` (https://www.gapminder.org/data/ indicateur gdp `life expectancy`) qui donne l'espérance de vie dans les pays du monde au cours des deux derniers siècles (ainsi que des prédictions jusqu'en 2100)

Le code suivant permet le chargement et formatage du fichier.

```
X<-read.csv('lex.csv')
nom.de.pays<-X[,1]
rownames(X)<-nom.de.pays
X<-X[,-1]
X['Afghanistan',]; # Donne les esperances de vie pour l'Afghanistan
```

```
##             X1800 X1801 X1802 X1803 X1804 X1805 X1806 X1807 X1808 X1809 X1810
## Afghanistan  28.2  28.2  28.2  28.2  28.2  28.2  28.1  28.1  28.1  28.1  28.1
##             X1811 X1812 X1813 X1814 X1815 X1816 X1817 X1818 X1819 X1820 X1821
## Afghanistan  28.1  28.1  28.1  28.1  28.1  28.1    28    28    28    28    28
##             X1822 X1823 X1824 X1825 X1826 X1827 X1828 X1829 X1830 X1831 X1832
## Afghanistan    28    28    28  27.9  27.9  27.9  27.9  27.9  27.9  27.9  27.9
##             X1833 X1834 X1835 X1836 X1837 X1838 X1839 X1840 X1841 X1842 X1843
## Afghanistan  27.9  27.9  27.9  27.8  27.8  27.8  27.8  27.8  27.8  27.8  27.8
##             X1844 X1845 X1846 X1847 X1848 X1849 X1850 X1851 X1852 X1853 X1854
## Afghanistan  27.8  27.8  27.7  27.7  27.7  27.7  27.7  27.9    28  28.2  28.3
##             X1855 X1856 X1857 X1858 X1859 X1860 X1861 X1862 X1863 X1864 X1865
## Afghanistan  28.5  28.6  28.8  28.9  29.1  29.2  29.4  29.5  29.7  29.8    30
##             X1866 X1867 X1868 X1869 X1870 X1871 X1872 X1873 X1874 X1875 X1876
## Afghanistan  30.1  30.3  30.4  30.6  30.7  30.9    31  31.2  31.3  31.5  31.6
##             X1877 X1878 X1879 X1880 X1881 X1882 X1883 X1884 X1885 X1886 X1887
## Afghanistan  31.8  31.9  32.1  32.3  32.4  32.5  32.7  32.9    33  33.1  33.3
##             X1888 X1889 X1890 X1891 X1892 X1893 X1894 X1895 X1896 X1897 X1898
## Afghanistan  33.4  33.6  33.7  33.9    34  34.2  34.3  34.5  34.6  34.8  34.9
##             X1899 X1900 X1901 X1902 X1903 X1904 X1905 X1906 X1907 X1908 X1909
## Afghanistan  35.1  35.2  35.4  35.5  35.7  35.8    36  36.1  36.2  36.4  36.5
##             X1910 X1911 X1912 X1913 X1914 X1915 X1916 X1917 X1918 X1919 X1920
## Afghanistan  36.7  36.8    37  37.1  37.3  37.4  37.6  37.7  9.88    38  38.1
##             X1921 X1922 X1923 X1924 X1925 X1926 X1927 X1928 X1929 X1930 X1931
## Afghanistan  38.3  38.4  38.6  38.7  38.9    39  39.2  39.3  39.4  39.6  39.7
##             X1932 X1933 X1934 X1935 X1936 X1937 X1938 X1939 X1940 X1941 X1942
## Afghanistan  39.9    40  40.2  40.3  40.5  40.6  40.7  40.9    41  41.2  41.4
##             X1943 X1944 X1945 X1946 X1947 X1948 X1949 X1950 X1951 X1952 X1953
## Afghanistan  41.5  41.7  41.9    42  42.2  42.4  42.5  42.7  42.9  43.1  43.5
##             X1954 X1955 X1956 X1957 X1958 X1959 X1960 X1961 X1962 X1963 X1964
```

```
## Afghanistan  43.3  43.9  44.1  44.3  44.5  44.7    45  45.3  45.5  45.7  45.9
##             X1965 X1966 X1967 X1968 X1969 X1970 X1971 X1972 X1973 X1974 X1975
## Afghanistan  46.1  46.3  46.5  46.7  46.9  47.1  47.3  47.3  47.3  47.4  47.5
##             X1976 X1977 X1978 X1979 X1980 X1981 X1982 X1983 X1984 X1985 X1986
## Afghanistan  47.7  47.9  46.4  44.7  43.7  44.3  44.1  42.3  39.9    42  43.3
##             X1987 X1988 X1989 X1990 X1991 X1992 X1993 X1994 X1995 X1996 X1997
## Afghanistan  45.9  48.5  52.7  53.8  53.8  54.2  54.4  53.9  54.3  54.7  54.5
##             X1998 X1999 X2000 X2001 X2002 X2003 X2004 X2005 X2006 X2007 X2008
## Afghanistan  53.3  54.7  54.7  54.8  55.5  56.5  57.1  57.6    58  58.5  59.2
##             X2009 X2010 X2011 X2012 X2013 X2014 X2015 X2016 X2017 X2018 X2019
## Afghanistan  59.9  60.5    61  61.4  61.9  61.9  61.9    62  62.9  62.7  63.3
##             X2020 X2021 X2022 X2023 X2024 X2025 X2026 X2027 X2028 X2029 X2030
## Afghanistan  62.3  61.8  62.6    64  64.8  65.1  65.4  65.6  65.9  66.1  66.3
##             X2031 X2032 X2033 X2034 X2035 X2036 X2037 X2038 X2039 X2040 X2041
## Afghanistan  66.6  66.8    67  67.2  67.4  67.6  67.8    68  68.2  68.3  68.5
##             X2042 X2043 X2044 X2045 X2046 X2047 X2048 X2049 X2050 X2051 X2052
## Afghanistan  68.7  68.9    69  69.2  69.4  69.5  69.7  69.8    70  70.2  70.3
##             X2053 X2054 X2055 X2056 X2057 X2058 X2059 X2060 X2061 X2062 X2063
## Afghanistan  70.5  70.6  70.8  70.9  71.1  71.2  71.3  71.5  71.6  71.8  71.9
##             X2064 X2065 X2066 X2067 X2068 X2069 X2070 X2071 X2072 X2073 X2074
## Afghanistan  72.1  72.2  72.3  72.5  72.6  72.8  72.9    73  73.2  73.3  73.5
##             X2075 X2076 X2077 X2078 X2079 X2080 X2081 X2082 X2083 X2084 X2085
## Afghanistan  73.6  73.8  73.9  74.1  74.2  74.3  74.5  74.6  74.8  74.9  75.1
##             X2086 X2087 X2088 X2089 X2090 X2091 X2092 X2093 X2094 X2095 X2096
## Afghanistan  75.2  75.4  75.5  75.6  75.8  75.9  76.1  76.2  76.4  76.5  76.7
##             X2097 X2098 X2099 X2100
## Afghanistan  76.8    77  77.1  77.3
```

```
X[,'X2024'];       # Donne les esperances de vie pour l'annee 2024
```

```
##   [1] 64.8 66.1 79.5 83.0 74.8 77.4 76.7 77.2 83.7 83.0 71.6 64.8 82.2 65.5 62.7
##  [16] 76.0 74.2 77.8 76.9 77.8 75.0 75.3 73.2 76.9 76.9 74.2 74.6 63.0 53.4 83.0
##  [31] 84.7 81.2 78.6 65.3 64.3 66.0 66.2 81.2 70.0 74.8 81.1 79.6 81.6 80.2 82.0
##  [46] 68.1 73.1 81.9 74.2 77.3 77.3 71.8 65.2 83.7 78.8 70.4 82.7 69.1 83.6 64.9
##  [61] 68.5 81.8 74.3 67.2 62.3 67.9 62.0 66.8 81.9 74.1 73.5 68.1  0.0 72.8 79.5
##  [76] 64.9 77.4 72.1 72.1 82.8 78.9 73.9 84.8 83.7 83.9 77.0 79.0 85.5 72.5 67.7
##  [91] 74.2 70.8 61.5 72.9 83.5 82.4 70.4 77.3 67.4 77.0 76.0  0.0 78.3 52.6 77.0
## [106] 83.7 76.5 74.5 80.6 75.1 66.8 80.2 76.7 66.2 75.2 63.0 83.3 70.6 76.8 69.4
## [121] 59.8 72.0 76.2 66.0 75.9 66.3 63.8 65.8 76.3 82.4 83.7 72.6 64.3 82.5 75.1
## [136] 66.6 81.0 81.3 72.4 68.5 65.5 78.9 73.8 82.6 77.2 75.7 77.2 76.3 73.8 69.8
## [151] 75.5 71.1 70.1 85.5 59.7 62.8 76.8 82.9 59.6 76.5 65.1 71.4 73.4 78.4 82.3
## [166] 83.6 59.3 74.6 74.7 61.4 66.2 79.6 70.3 71.5 71.9 73.7 75.7 78.9 79.5 69.0
## [181] 81.3 68.3 67.3 74.7 78.3 79.7 69.0 73.8 76.0 75.3 66.4 71.2 69.0 65.8 64.0
## [196] 61.8
```

a. Considérer l'année 2024. Calculer la moyenne, la moyenne tronquée, la médiane. Que remarquer vous
b. Dessiner un boxplot ? Commenter.
c. Pensez vous que l'année 2024 contienne des ou une données aberrantes ? Justifier.
d. Si oui recommencer les analyses précédentes sans la ou les données qui vous semblent aberrantes.

2. Comparer les années 1824, 1924 et 2024 à l'aide de représentations graphiques.
3. Quels pays possèdent l'espérance de vie la plus petite, la plus grande.
4. Quels sont les quartiles ?
5. Ecrire un code qui forme 4 groupes de pays suivant les quartiles de l'espérance de vie en 2024. Créer une représentation graphique qui illustre ces groupes.
6. Dans quel groupe aurait été classée la France de l'année 1824 ?

Exercice 2

1. Charger les jeu de données `gdp_pcap.csv` (https://www.gapminder.org/data/ onglet indicateur `gdp per capita`) qui donne le PIB par habitant en mesurant la valeur de tout ce qui est produit dans un pays pendant un an, divisé par le nombre de personnes. L'unité est en dollars constants ajustés pour l'inflation aux prix de 2017. Comme le coût de la vie varie d'un pays à l'autre, nous utilisons une monnaie appelée "dollars internationaux", qui est ajustée en fonction de la Parité de Pouvoir d'Achat (PPA). Il s'agit d'une monnaie virtuelle qui permet de meilleures comparaisons. Un tel dollar achèterait dans chaque pays une quantité comparable de biens et services à ce qu'un dollar américain achèterait aux États-Unis. Le PIB par habitant est le PIB divisé par la population du pays, ce qui donne une estimation approximative du revenu annuel moyen des citoyens.

Le code suivant permet le chargement et formatage du fichier.

```
X<-read.csv("gdp_pcap.csv")
nom.de.pays<-X[,1]
rownames(X)<-nom.de.pays
X<-X[,-1]

convert_k_to_number <- function(string) {
  # Vérifie si la chaîne se termine par 'k' ou 'K'
  if (grepl("k$", tolower(string))) {
    # Supprime la lettre 'k', remplace le point décimal par rien et convertit en nombre
    num <- as.numeric(gsub("k$", "", tolower(string))) * 1000
  } else {
    # Convertit directement en nombre si 'k' n'est pas présent
    num <- as.numeric(string)
  }
  return(num)
}

for (i in 1:nrow(X))
  for (j in 1:ncol(X))
    if (is.character(X[i,j])) {X[i,j]<-convert_k_to_number(X[i,j])}
```

   a. Tracer pour tout les pays du jeu de données un graphe qui croise pib par habitant et espérance de vie pour les années 1900 et 2000.
   b. Quand dit on que deux variables aléatoires sont indépendantes ?
   c. D'après vos graphiques de la question a), est-ce que `pib par habitant` et `espérance de vie` sont indépendants ?
   d. Former quatre groupes de pays suivant leur `pib par habitant` et illustrer (comme dans la question 5 de l'exercice 1).

Exercice 3

Soit le tableau suivant décrivant les relations entre deux variables aléatoires X et Y

	$Y = m_1$	$Y = m_2$	$Y = m_3$	$Y = m_4$
$X = l_1$	n_{11}	n_{12}	n_{13}	n_{14}
$X = l_2$	n_{21}	n_{22}	n_{23}	n_{24}
$X = l_3$	n_{31}	n_{32}	n_{33}	n_{34}
$X = l_4$	n_{41}	n_{42}	n_{43}	n_{44}

Nous noterons

- $n_{i\bullet} = \sum_j n_{ij}$ (marge en ligne)
- $n_{\bullet j} = \sum_i n_{ij}$ (marge en colonne)
- $n = \sum_{ij} n_{ij} = \sum_i n_{i\bullet} = \sum_j n_{\bullet j}$, le nombre total d'individus de l'échantillon.

Partie 1

1. Comment s'appelle ce type de tableau ?
2. Donner une estimation de la probabilité $P(X = l_1, Y = m_2)$.
3. Donner une estimation de la probabilité $P(Y = m_2)$.
4. Donner une estimation de la probabilité $P(X = l_1 | Y = m_2)$
5. Si les deux variables X et Y étaient indépendantes comment pourrait on estimer la probabilité jointe $P(X = l_i, Y = m_j)$? $P(X = l_i, Y = m_j) = P(X = l_i)P(Y = m_j)$

Partie 2

1. Créer un tableau où chaque ligne est un pays, qui possède deux colonnes: la première donne le numéro du groupe de l'espérance de vie du pays (1, 2 , 3 ou 4), la seconde le numéro du groupe du PIB par habitant (1, 2, 3 ou 4).
2. A partir de ce tableau créer un tableau similaire à celui décrit en partie 1 où X est le groupe d'espérance de vie et Y le groupe de PIB par habitant.
3. Calculer le tableau théorique que vous auriez obtenu si les variables X et Y étaient indépendantes.