

Examen — STT-1100 : Introduction à la science des données

Mission ONU : développement humain à l'ère numérique

Présentation du jeu de données

Le fichier `gapminder_rich.csv` contient des indicateurs socio-économiques et démographiques par pays entre **1960 et 2016**.

Chaque ligne représente un pays pour une année donnée.

Variable	Description	Type
<code>country</code>	Nom du pays	caractère
<code>continent</code>	Continent	facteur
<code>year</code>	Année d'observation	numérique
<code>life_expectancy</code>	Espérance de vie à la naissance (années)	numérique
<code>fertility</code>	Nombre moyen d'enfants par femme	numérique
<code>population</code>	Nombre d'habitants dans le pays	numérique
<code>gdp</code>	PIB par habitant (USD courants)	numérique
<code>net_users</code>	Proportion d'utilisateurs d'Internet (classes : >0%, >5%, >15%, >35%, >60%)	facteur ordonné

Mise en situation — Mission ONU

Tu travailles comme **assistant · e en science des données** pour le **Programme des Nations Unies pour le Développement (PNUD)**, basé à Montréal.

Ta superviseure, **Mme Aïcha Diallo**, prépare le prochain *Rapport mondial sur le développement humain*. Elle te confie la mission suivante :

« Je veux un portrait clair du monde en 2010. Comment la prospérité, la fécondité et la connectivité numérique sont-elles reliées à la qualité de vie ?

Prépare un rapport visuel et concis, facile à intégrer dans notre publication. »

Tu dois donc produire un **document Quarto** qui répond à des questions précises, chacune représentant une étape de ton mandat.

Le rapport doit être **exécutable sans erreur**, clair, commenté, et accompagné d'interprétations brèves et professionnelles.

Consignes générales

- Remets un seul fichier `.qmd` qui s'exécute sans erreur et produit un `.html`.
 - Structure ton rapport selon les sections ci-dessous.
 - Chaque graphique doit comporter un titre, des axes clairs et une brève interprétation.
 - Utilise un ton professionnel (rédige comme si tu préparais un livrable pour le PNUD).
 - Aucune connexion Internet n'est permise, mais tu peux utiliser tes documents papier.
-

1. Préparer et comprendre les données (15 pts)

Objectif : vérifier la qualité des données que tu t'apprêtes à utiliser pour les analyses du rapport ONU.

- 1.1. Importe le fichier `gapminder_rich.csv` dans un tibble `gap` et affiche un aperçu (`glimpse`). (3 pts)
 - 1.2. Indique les années minimale et maximale disponibles, ainsi que le nombre total de pays différents. (3 pts)
 - 1.3. Crée un sous-ensemble `gap2010` contenant uniquement l'année 2010. (3 pts)
 - 1.4. Vérifie la présence de valeurs manquantes pour `life_expectancy`, `fertility` et `gdp`. (3 pts)
 - 1.5. En une ou deux phrases, explique pourquoi l'année 2010 constitue un bon choix pour le rapport. (3 pts)
-

2. Étudier la relation entre santé, fécondité et richesse (25 pts)

Objectif : produire des statistiques et un graphique illustrant les différences mondiales de conditions de vie.

- 2.1. Calcule la moyenne et la médiane de `life_expectancy`, `fertility` et `gdp` en 2010. (5 pts)
 - 2.2. Crée un **nuage de points** représentant `fertility` (x) et `life_expectancy` (y), coloré par `continent`. (7 pts)
 - 2.3. Crée un **nuage de points** représentant `fertility` (x) et `gdp` (y), coloré par `continent` avec la taille des points selon la variable `life_expectancy`. (7 pts)
 - 2.4. Rédige un court paragraphe (3 à 4 phrases) expliquant la tendance observée et les différences entre continents. (6 pts)
-

3. Analyser la distribution de l'espérance de vie (15 pts)

Objectif : décrire la dispersion de l'espérance de vie dans le monde pour le rapport.

- 3.1. Trace un **histogramme** de `life_expectancy` en 2010.
Ajoute des lignes verticales pour la moyenne et la médiane. (6 pts)
 - 3.2. Décris en 2 à 3 phrases la forme de la distribution : le centre, la dispersion, et la symétrie (ou non). (4 pts)
 - 3.3. Identifie les pays avec la plus faible et la plus haute espérance de vie en 2010. Rédige une phrase résumant ces extrêmes. (2 pts)
 - 3.4. Dans un même graphique, trace des **diagrammes en boîtes** de `life_expectancy` par `continent`. Résume brièvement les différences entre continents (2 à 3 phrases). (3 pts)
-

4. Connectivité numérique et espérance de vie (20 pts)

Objectif : étudier si les pays plus connectés affichent une meilleure qualité de vie.

- 4.1. Transforme la variable `net_users` en **facteur ordonné** du plus faible au plus fort. (4 pts)
- 4.2. Calcule la moyenne de `life_expectancy` pour chaque classe de `net_users`. (5 pts)
- 4.3. Crée un **diagramme en boîtes** de `life_expectancy` selon les classes de `net_users`. (6 pts)
- 4.4. Interprète les résultats numériques et graphiques (3 à 5 phrases) : observe-t-on un gradient de qualité de vie entre les classes ? (5 pts)

5. Graphique de synthèse pour le rapport (25 pts)

Objectif : réaliser une figure finale complète (type *scatter multi-esthétiques*) pour le rapport et en donner une interprétation claire.

5.1 Crée un nuage de points où :

- **x** = `log(gdp)` (utilise une échelle logarithmique)
 - **y** = `life_expectancy`
 - **couleur** = `continent`
 - **forme** = `net_users` (facteur ordonné)
 - **taille** = `population`
- (10 pts)

5.2 Ajoute un **titre informatif**, des **étiquettes d'axes** claires, des **légendes lisibles** (population, continent, net_users).

Ajuste la **transparence (alpha)** et/ou l'échelle des tailles pour limiter le chevauchement des points. (7 pts)

5.3 Rédige un **paragraphe d'interprétation (80–120 mots)** exposant le message clé destiné au rapport

(tendance générale, différences entre continents, rôle de la connectivité et du poids démographique, prudence interprétative). (8 pts)

Total : 100 points