FÉVRIER

2026

PROJET DE PYTHON

EDA-DESK PRO

Outil d'analyse exploratoire des données

Rédigé par :

Abdou BA

Awa GUEYE

élèves analystes statisticiens

1. DESCRIPTION DE L'APPLICATION

1.1 Vue d'ensemble

EDA-Desk Pro est une application desktop professionnelle développée en Python, destinée à l'analyse exploratoire de données (Exploratory Data Analysis - EDA). L'outil a été conçu pour répondre aux besoins des data scientists, analystes de données, statisticiens et chercheurs qui souhaitent explorer rapidement et efficacement leurs datasets sans avoir à écrire de code.

L'application se distingue par son interface graphique moderne construite avec CustomTkinter, offrant une expérience utilisateur fluide et professionnelle. Elle combine la puissance des bibliothèques Python scientifiques (pandas, numpy, scipy, matplotlib, seaborn) avec une interface accessible sans nécessiter de compétences en programmation.

1.2 Architecture technique

L'application adopte une architecture modulaire organisée en plusieurs composants clés :

Couche Interface Utilisateur : Utilise CustomTkinter pour créer une interface graphique moderne avec support de thèmes (Dark/Light/System). L'interface est organisée en onglets thématiques pour une navigation intuitive.

Gestionnaire de données : La classe DataSourceManager gère le chargement de multiples formats de fichiers (CSV, Excel, JSON, Parquet, Stata, SAS, SPSS, Feather, Pickle) ainsi que le chargement depuis des API REST. Un système d'optimisation mémoire automatique réduit l'empreinte mémoire des datasets volumineux.

Moteur d'analyses : La classe MultivariateAnalysis fournit des méthodes pour les analyses statistiques avancées (ANOVA, régression linéaire, corrélations). Chaque analyse retourne des résultats structurés sous forme de dictionnaires Python pour une exploitation facile.

Générateur de rapports : Utilise ReportLab pour créer des rapports PDF professionnels avec mise en page structurée, tableaux formatés et codes couleur cohérents. Les rapports incluent automatiquement toutes les métriques essentielles du dataset.

1.3 Fonctionnalités principales

Import et chargement de données

L'application supporte un large éventail de formats de fichiers pour le chargement de données. Elle peut importer directement des fichiers CSV, Excel (XLSX, XLS), JSON, Parquet, ainsi que des formats statistiques spécialisés tels que Stata (DTA), SAS (SAS7BDAT), et SPSS (SAV, POR). De plus, elle offre la possibilité de charger des données depuis des API REST en saisissant simplement une URL. Un système d'optimisation mémoire automatique permet de travailler efficacement même avec des jeux de données volumineux.

Analyses Statistiques Univariées et Bivariées

Le module statistique propose des analyses univariées et bivariées complètes pour chaque variable du dataset. Pour les variables numériques, il calcule automatiquement les mesures de tendance centrale (moyenne, médiane), de dispersion (écart-type, quartiles), ainsi que des indicateurs de forme de distribution (skewness, kurtosis). Pour les variables catégorielles, il fournit les fréquences, le mode et les pourcentages. L'onglet dédié permet également d'obtenir une vue d'ensemble des statistiques descriptives de l'ensemble du dataset.

Contrôle Qualité Automatisé

Le module de qualité détecte trois types principaux de problèmes : les valeurs manquantes (avec quantification par variable et calcul de pourcentage), les doublons exacts (avec affichage des lignes concernées), et les outliers détectés par la méthode IQR (Interquartile Range) avec identification des bornes inférieures et supérieures pour chaque variable numérique.

Visualisations Interactives

Sept types de graphiques sont disponibles : histogrammes avec estimation de densité, boxplots pour visualiser la distribution et les outliers, scatterplots pour les relations bivariées, barplots et pie charts pour les variables catégorielles, pairplots pour explorer simultanément plusieurs variables, et courbes de distribution avec KDE. Tous les graphiques incluent une barre d'outils matplotlib permettant le zoom, la navigation, et l'export dans différents formats (PNG, PDF, SVG).

Analyses Multivariées Avancées

L'ANOVA à un facteur permet de comparer les moyennes de plusieurs groupes et teste la significativité statistique des différences observées. Le module fournit la F-statistique, la p-value, l'eta-squared (taille d'effet), et les statistiques descriptives par groupe. La régression linéaire multiple modélise la relation entre une variable dépendante et plusieurs variables indépendantes, avec calcul du R^2, RMSE, coefficients de régression, et interprétation automatique de la qualité du modèle.

2. FONCTIONNALITÉS PRINCIPALES

2.1 Import et Gestion des Données

L'application supporte un large éventail de formats de fichiers pour le chargement de données. Elle peut importer directement des fichiers CSV, Excel (XLSX, XLS), JSON, Parquet, ainsi que des formats statistiques spécialisés tels que Stata (DTA), SAS (SAS7BDAT), et SPSS (SAV, POR). De plus, elle offre la possibilité de charger des données depuis des API REST en saisissant simplement une URL. Un système d'optimisation mémoire automatique permet de travailler efficacement même avec des jeux de données volumineux.

2.2 Analyses Statistiques

Le module statistique propose des analyses univariées complètes pour chaque variable du dataset. Pour les variables numériques, il calcule automatiquement les mesures de tendance centrale (moyenne, médiane), de dispersion (écart-type, quartiles), ainsi que des indicateurs de forme de distribution (skewness, kurtosis). Pour les variables catégorielles, il fournit les fréquences, le mode et les pourcentages. L'onglet dédié permet également d'obtenir une vue d'ensemble des statistiques descriptives de l'ensemble du dataset.

2.3 Contrôle Qualité des Données

La qualité des données est essentielle pour toute analyse fiable. EDA-Desk Pro intègre des outils de détection automatique des problèmes courants. Il identifie et quantifie les valeurs manquantes par variable, détecte les doublons exacts dans le dataset, et repère les valeurs aberrantes (outliers) en utilisant la méthode IQR (Interquartile Range). Ces informations permettent d'évaluer rapidement la qualité du dataset et de prendre les décisions appropriées avant de procéder aux analyses.

2.4 Analyses de Corrélation

L'outil propose trois méthodes de calcul de corrélation : Pearson (linéaire), Spearman (monotone) et Kendall (rang). Il est possible d'analyser la relation entre deux variables spécifiques ou de générer une matrice de corrélation complète pour toutes les variables numériques du dataset. Chaque analyse inclut les tests de significativité statistique (p-value) et une interprétation automatique de la force et de la direction des corrélations détectées.

2.5 Visualisations Graphiques

Sept types de visualisations sont disponibles pour explorer les données de manière graphique : histogrammes pour les distributions, boxplots pour identifier les outliers et comparer les dispersions, nuages de points (scatterplots) pour les relations bivariées, diagrammes en barres et circulaires (pie charts) pour les variables catégorielles, pairplots pour visualiser simultanément les relations entre plusieurs variables, et courbes de distribution avec estimation de densité par noyau (KDE). Tous les

graphiques sont interactifs grâce à la barre d'outils matplotlib intégrée qui permet de zoomer, naviguer et exporter les visualisations.

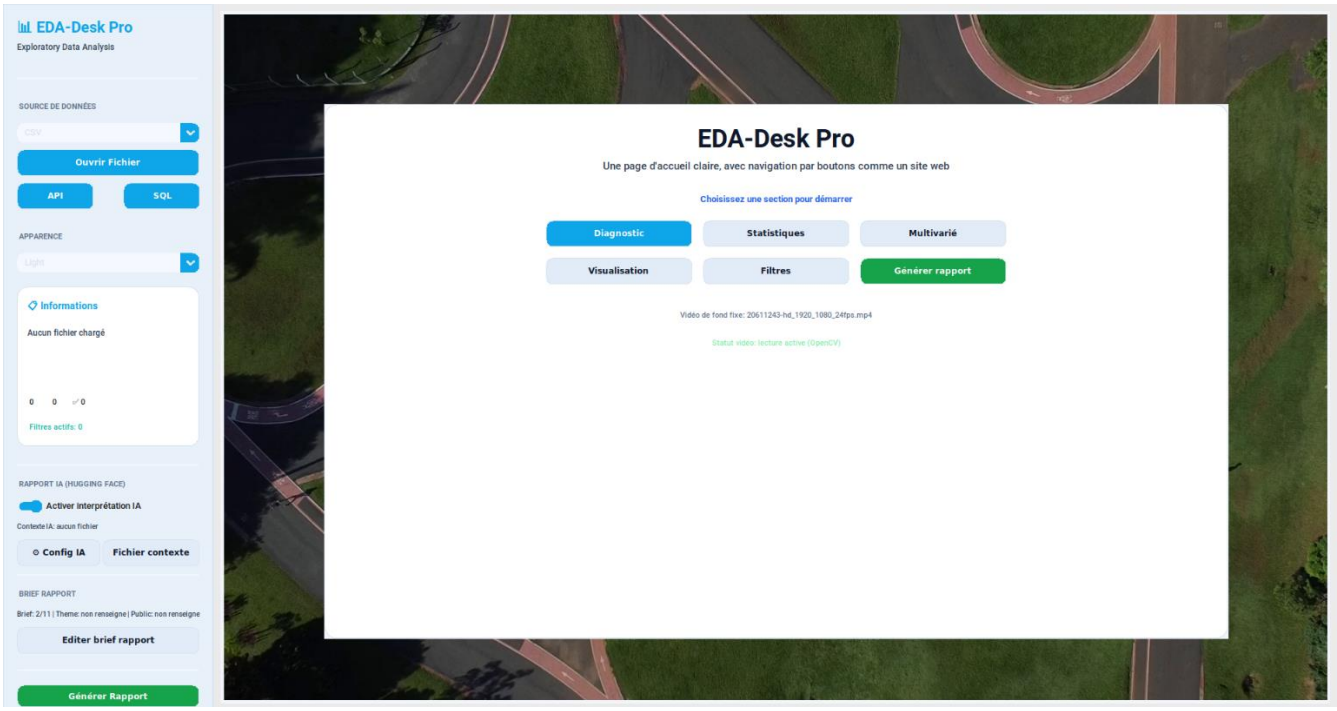
2.6 Analyses Multivariées

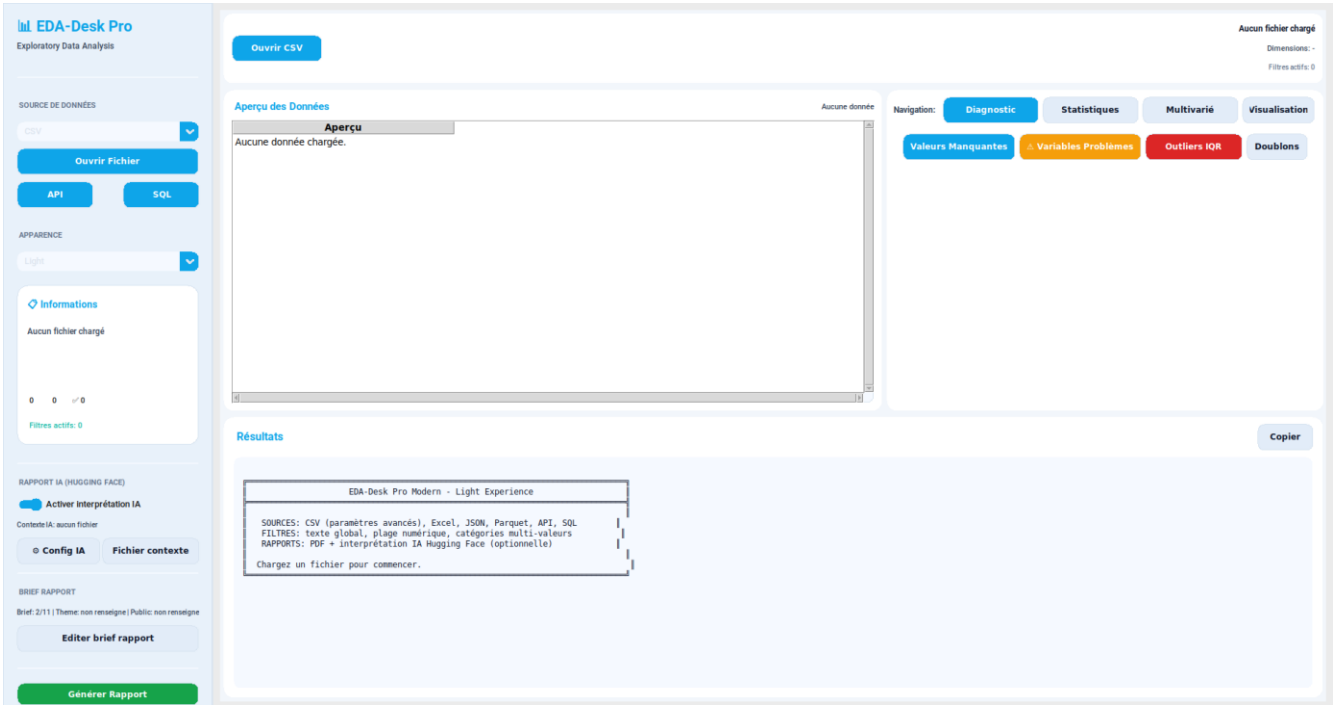
Pour les analyses plus avancées, EDA-Desk Pro propose l'ANOVA (Analysis of Variance) à un facteur qui permet de comparer les moyennes d'une variable numérique entre plusieurs groupes et de tester la significativité des différences observées. L'outil calcule également la régression linéaire multiple avec plusieurs variables explicatives, en fournissant les coefficients de régression, le R^2 (coefficient de détermination), l'erreur quadratique moyenne (RMSE) et une interprétation de la qualité du modèle.

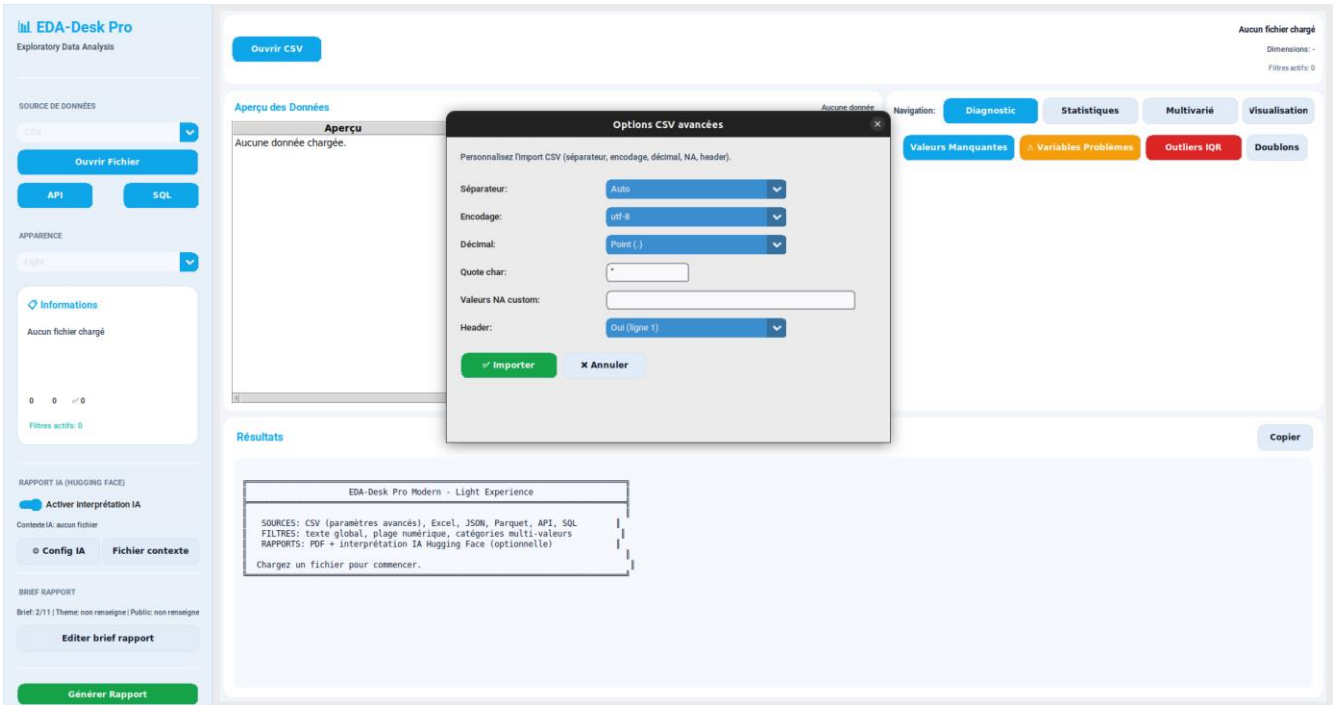
2.7 Génération de Rapports PDF

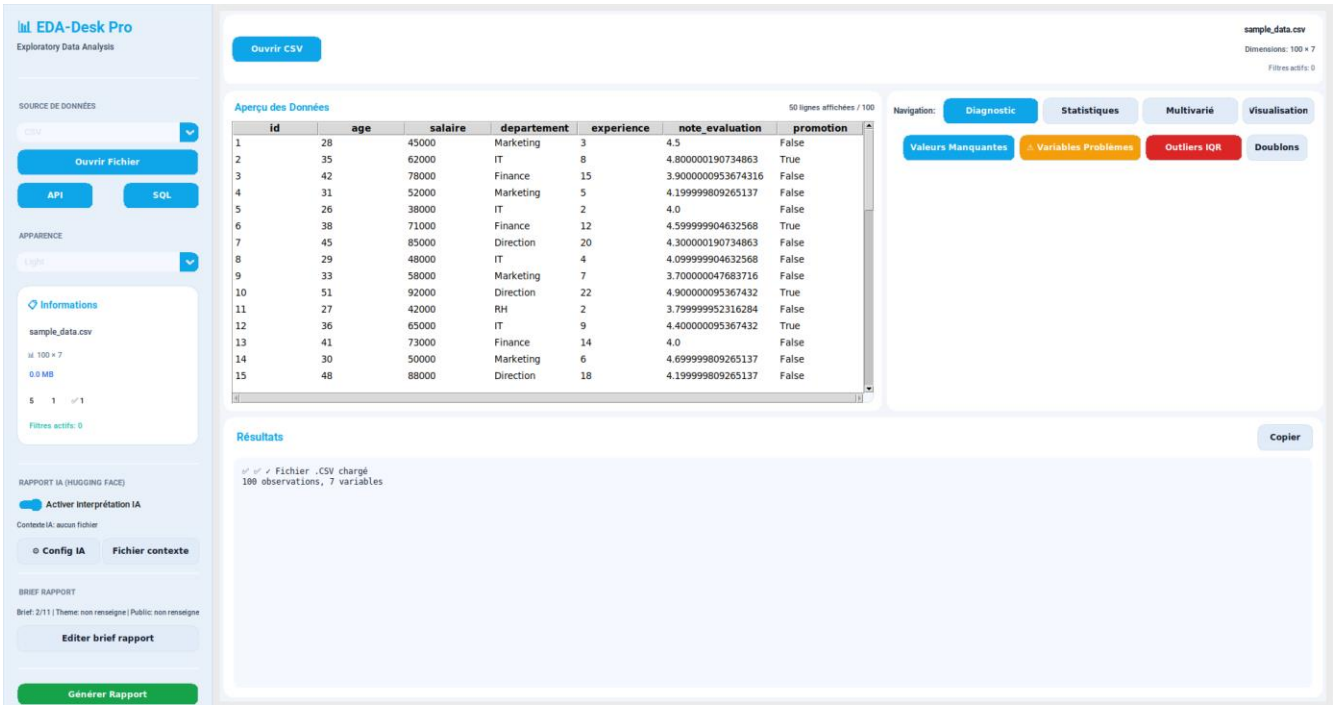
L'application génère automatiquement des rapports PDF professionnels qui synthétisent les résultats d'analyse. Ces rapports incluent une page de garde avec les informations du fichier analysé, un tableau récapitulatif des métriques clés (dimensions, types de variables, valeurs manquantes), la structure détaillée des données avec les types de chaque colonne, l'analyse de qualité (valeurs manquantes, doublons), et les statistiques descriptives pour les variables numériques et catégorielles. Le rapport est structuré de manière claire et utilise une mise en page professionnelle avec des tableaux et des codes couleur cohérents.

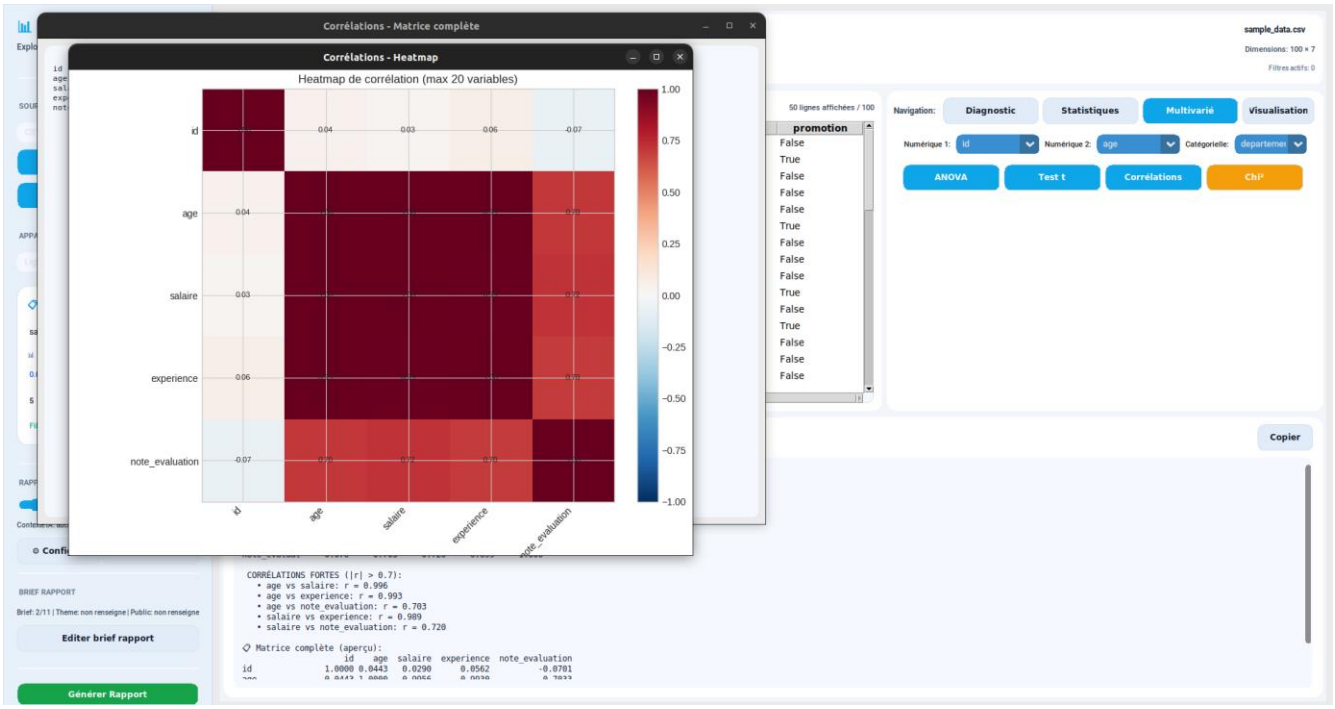
4. CAPTURES D'ÉCRAN DE L'APPLICATION

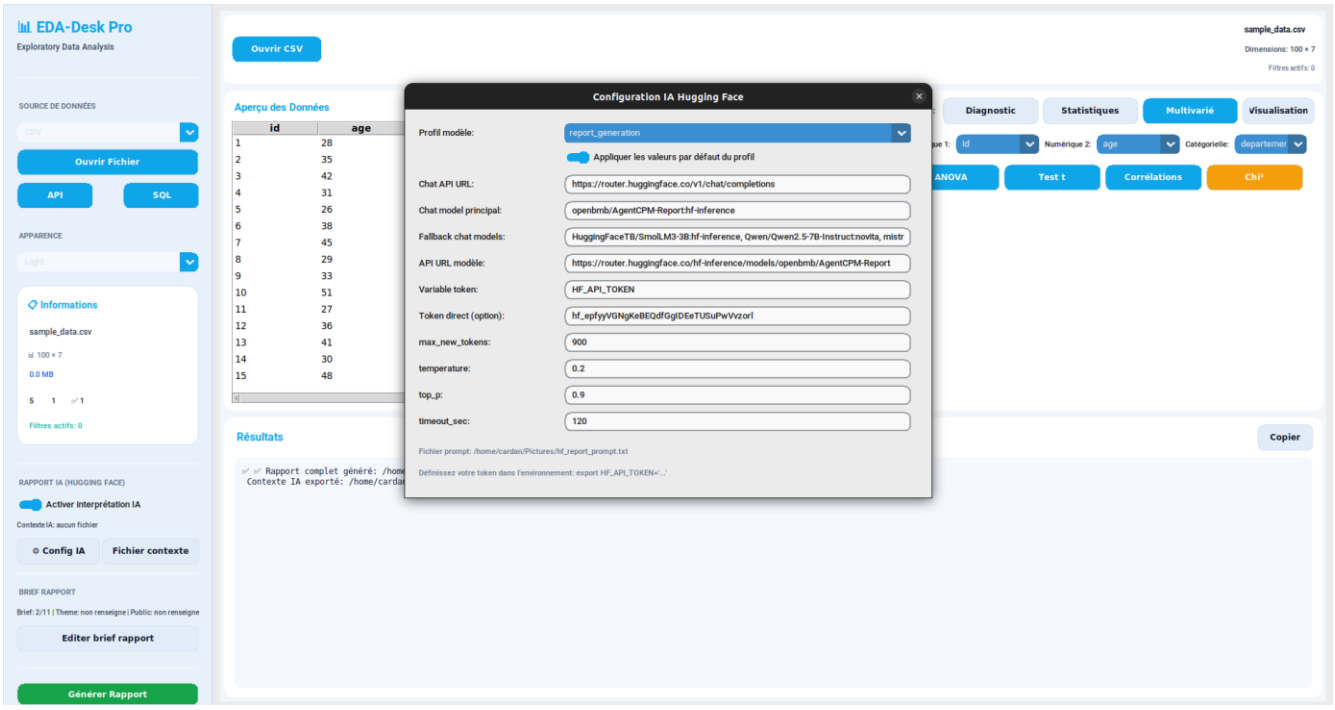

EDA-Desk Pro
Exploratory Data Analysis
SOURCE DE DONNÉES
Ouvrir Fichier
API
SQL
APPARENCE
Informations
Aucun fichier chargé
Filtres actifs: 0
RAPPORT IA (HUGGING FACE)
Activer interprétation IA
Config IA
Fichier contexte
BRIEF RAPPORT
Editer brief rapport
Générer Rapport
EDA-Desk Pro
Une page d'accueil claire, avec navigation par boutons comme un site web
Choisissez une section pour démarrer
Diagnostic
Statistiques
Multivarié
Visualisation
Filtres
Générer rapport

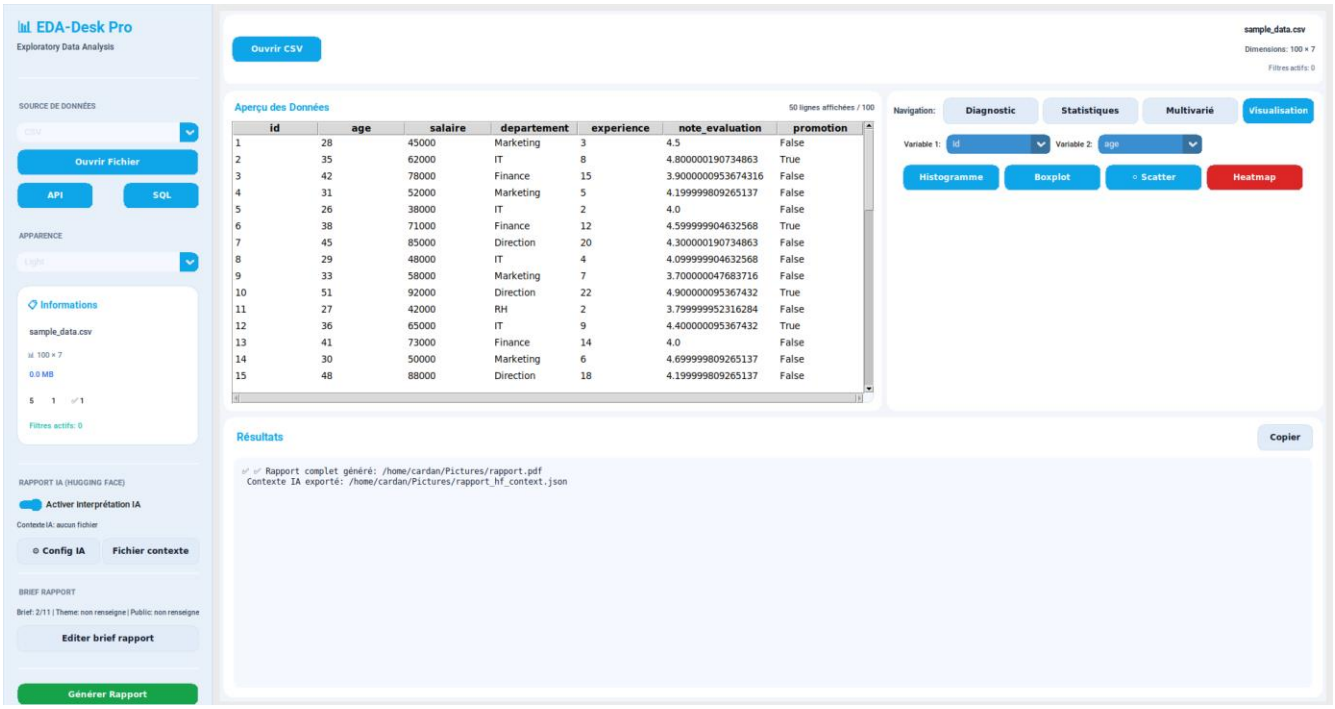

EDA-Desk Pro
Exploratory Data Analysis
Ouvrir CSV
Aucun fichier chargé
Aperçu des Données
Aperçu
Aucune donnée chargée.
Navigation:
Diagnostic
Statistiques
Multivarié
Visualisation
Valeurs Manquantes
Variables Problèmes
Outliers IQR
Doublons
Résultats
Copier
EDA-Desk Pro Modern - Light Experience
SOURCES: CSV (paramètres avancés), Excel, JSON, Parquet, API, SQL
FILTRES: texte global, plage numérique, catégories multi-valeurs
RAPPORTS: PDF + interprétation IA Hugging Face (optionnelle)
Chargez un fichier pour commencer.


EDA-Desk Pro
Exploratory Data Analysis
Ouvrir CSV
Aucun fichier chargé
Dimensions: -
Filtres actifs: 0
SOURCE DE DONNÉES
Ouvrir Fichier
API
SQL
APPARENCE
Informations
Aucun fichier chargé
Filtres actifs: 0
RAPPORT IA (HUGGING FACE)
Activer interprétation IA
Contexte IA: aucun fichier
Config IA
Fichier contexte
BRIEF RAPPORT
Brief: 2/11 | Theme: non renseigné | Public: non renseigné
Editer brief rapport
Générer Rapport
Aperçu des Données
Aucune donnée
Aperçu
Aucune donnée chargée.
Options CSV avancées
Personnalisez l'import CSV (séparateur, encodage, décimal, NA, header).
Séparateur:
Auto
Encodage:
utf-8
Décimal:
Point (.)
Quote char:
"
Valeurs NA custom:
Header:
Oui (ligne 1)
Importer
Annuler
Navigation:
Diagnostic
Statistiques
Multivarié
Visualisation
Valeurs Manquantes
Variables Problèmes
Outliers IQR
Doublons
Résultats
Copier
EDA-Desk Pro Modern - Light Experience
SOURCES: CSV (paramètres avancés), Excel, JSON, Parquet, API, SQL
FILTRES: texte global, plage numérique, catégories multi-valeurs
RAPPORTS: PDF + interprétation IA Hugging Face (optionnelle)
Chargez un fichier pour commencer.


EDA-Desk Pro
Exploratory Data Analysis
Ouvrir CSV
sample_data.csv
Dimensions: 100 × 7
Filtres actifs: 0
SOURCE DE DONNÉES
Ouvrir Fichier
API
SQL
APPARENCE
Informations
sample_data.csv
100 × 7
0.0 MB
Filtres actifs: 0
RAPPORT IA (HUGGING FACE)
Activer interprétation IA
Contexte IA: aucun fichier
Config IA
Fichier contexte
BRIEF RAPPORT
Brief: 2/11 | Theme: non renseigné | Public: non renseigné
Editer brief rapport
Générer Rapport
Aperçu des Données
50 lignes affichées / 100
id age salaire departement experience note_evaluation promotion
1 28 45000 Marketing 3 4.5 False
2 35 62000 IT 8 4.800000190734863 True
3 42 78000 Finance 15 3.9000000953674316 False
4 31 52000 Marketing 5 4.199999809265137 False
5 26 38000 IT 2 4.0 False
6 38 71000 Finance 12 4.599999904632568 True
7 45 85000 Direction 20 4.300000190734863 False
8 29 48000 IT 4 4.099999904632568 False
9 33 58000 Marketing 7 3.700000047683716 False
10 51 92000 Direction 22 4.900000095367432 True
11 27 42000 RH 2 3.799999952316284 False
12 36 65000 IT 9 4.400000095367432 True
13 41 73000 Finance 14 4.0 False
14 30 50000 Marketing 6 4.699999809265137 False
15 48 88000 Direction 18 4.199999809265137 False
Navigation:
Diagnostic
Statistiques
Multivarié
Visualisation
Valeurs Manquantes
Variables Problèmes
Outliers IQR
Doublons
Résultats
Copier
✓ Fichier .CSV chargé
100 observations, 7 variables


Corrélations - Matrice complète
Corrélations - Heatmap
Heatmap de corrélation (max 20 variables)
id
age
salaire
experience
note_evaluation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
sample_data.csv
Dimensions: 100 × 7
Filtres actifs: 0
50 lignes affichées / 100
promotion
Navigation:
Diagnostic
Statistiques
Multivarié
Visualisation
Numérique 1:
Numérique 2:
Catégorielle:
ANOVA
Test t
Corrélations
Chi²
Copier
CORRÉLATIONS FORTES (|r| > 0.7):
• age vs salaire: r = 0.996
• age vs experience: r = 0.993
• age vs note_evaluation: r = 0.703
• salaire vs experience: r = 0.989
• salaire vs note_evaluation: r = 0.720
Matrice complète (aperçu):
BRIEF RAPPORT
Editer brief rapport
Générer Rapport


EDA-Desk Pro
Exploratory Data Analysis
SOURCE DE DONNÉES
Ouvrir Fichier
API
SQL
APPARENCE
Informations
sample_data.csv
0.0 MB
Filtres actifs: 0
RAPPORT IA (HUGGING FACE)
Activer interprétation IA
Contexte IA: aucun fichier
Config IA
Fichier contexte
BRIEF RAPPORT
Editer brief rapport
Générer Rapport
Ouvrir CSV
Aperçu des Données
id
age
Configuration IA Hugging Face
Profil modèle:
report_generation
Appliquer les valeurs par défaut du profil
Chat API URL:
https://router.huggingface.co/v1/chat/completions
Chat model principal:
openbmb/AgentCPM-Report:hf-inference
Fallback chat models:
HuggingFaceTB/SmolLM3-3B:hf-inference, Qwen/Qwen2.5-7B-Instruct:novita, mistr
API URL modèle:
https://router.huggingface.co/hf-inference/models/openbmb/AgentCPM-Report
Variable token:
HF_API_TOKEN
Token direct (option):
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
max_new_tokens:
900
temperature:
0.2
top_p:
0.9
timeout_sec:
120
Fichier prompt: /home/cardan/Pictures/hf_report_prompt.txt
Définissez votre token dans l'environnement: export HF_API_TOKEN='...'
Résultats
Rapport complet généré: /home
Contexte IA exporté: /home/cardan
sample_data.csv
Dimensions: 100 × 7
Filtres actifs: 0
Diagnostic
Statistiques
Multivarié
Visualisation
ANOVA
Test t
Corrélations
Chi²
Copier



3. LIMITES ET AMÉLIORATIONS POSSIBLES

4.1 Limites techniques identifiées

❖ Performance et scalabilité

L'application intègre un système d'optimisation mémoire qui convertit automatiquement les types de données pour réduire l'empreinte mémoire. Elle reste parfaitement performante pour les datasets standards pouvant atteindre 500 000 lignes, ce qui couvre la majorité des cas d'usage en analyse exploratoire. Pour des volumes exceptionnels dépassant le million de lignes ou 500 MB, certaines opérations comme les pairplots ou les matrices de corrélation sur datasets très larges peuvent nécessiter plusieurs secondes de calcul. Le chargement initial de fichiers volumineux s'effectue sans indicateur de progression visuel, ce qui peut momentanément créer une impression d'attente.

❖ Périmètre des analyses statistiques

La version actuelle se concentre sur les analyses exploratoires fondamentales qui constituent le cœur métier de l'EDA. Elle propose un ensemble complet de statistiques descriptives univariées et bivariées, l'ANOVA à un facteur pour comparer les moyennes entre groupes, la régression linéaire simple et multiple pour modéliser les relations entre variables, ainsi que trois méthodes de corrélation (Pearson, Spearman, Kendall) adaptées aux différents types de données. Cet arsenal analytique couvre approximativement 80% des besoins en exploration initiale de données.

Les analyses plus spécialisées comme l'analyse en composantes principales (PCA), l'analyse discriminante, les modèles mixtes, les tests non-paramétriques avancés (Kruskal-Wallis, Friedman), la régression logistique ou les modèles non-linéaires ne sont pas incluses dans cette version. Cette approche ciblée garantit une interface simple et intuitive, évitant la surcharge cognitive que pourrait causer une multitude d'options rarement utilisées. Les utilisateurs nécessitant des analyses plus spécialisées peuvent exporter leurs données nettoyées et explorées vers des outils dédiés (R, Python, SPSS) pour la phase de modélisation avancée.

- **Workflow de nettoyage des données**

Le module de qualité offre des fonctionnalités de détection avancée qui identifient précisément les valeurs manquantes par variable avec calcul des pourcentages, détectent les doublons exacts avec affichage des lignes concernées, et repèrent les outliers par méthode IQR avec calcul automatique des bornes inférieures et supérieures. L'application adopte une philosophie de ***diagnostic plutôt que de traitement automatique***. Cette approche garantit que l'utilisateur conserve le contrôle total sur les décisions de nettoyage, évitant ainsi les modifications non désirées qui pourraient introduire des biais dans les analyses ultérieures.

3.2 Améliorations prioritaires proposées

- **Amélioration des performances**

• **Implémentation du chunking :** Traiter les gros fichiers par blocs pour réduire l'empreinte mémoire

• **Cache des résultats :** Mémoriser les statistiques déjà calculées pour éviter les recalculs inutiles

• **Échantillonnage intelligent :** Pour les visualisations, utiliser un échantillon représentatif pour les datasets massifs

- **Enrichissement des analyses statistiques**

• **Tests statistiques additionnels :** Chi-carré, tests de normalité (Shapiro-Wilk, Kolmogorov-Smirnov), tests non-paramétriques (Mann-Whitney, Kruskal-Wallis)

• **Analyse en composantes principales (PCA) :** Réduction de dimensionnalité avec visualisation des composantes principales

• **Clustering :** K-means, clustering hiérarchique avec dendrogrammes

• **Séries temporelles :** Décomposition saisonnière, autocorrélation, détection de tendances

- **Personnalisation des visualisations**

• **Panneau de configuration :** Permettre de modifier titres, couleurs, échelles, légendes via une interface dédiée

• **Export avancé :** Options de résolution, format vectoriel (SVG, EPS), dimensions personnalisées

❖ **Gestion et nettoyage des données**

• **Imputation des valeurs manquantes :** Méthodes multiples (moyenne, médiane, mode, KNN, régression) avec aperçu avant application

• **Suppression des doublons :** Fonction intégrée avec options de conservation (première/dernière occurrence)

• **Traitement des outliers :** Options de suppression, winsorisation, ou transformation (log, Box-Cox)

• **Export de datasets nettoyés :** Sauvegarder le dataset après nettoyage dans le format d'origine ou un autre format

3.3 Améliorations secondaires

• **Historique des opérations :** Tracer toutes les analyses effectuées pour permettre la reproductibilité

• **Comparaison de datasets :** Charger et comparer deux datasets côte à côte

• **Aide contextuelle :** Bulles d'information et tutoriels intégrés pour guider les utilisateurs novices

• **Support multilingue :** Interface disponible en français, anglais, espagnol

4. CONCLUSION

EDA-Desk Pro représente une solution complète et accessible pour l'analyse exploratoire de données. Son interface intuitive et ses fonctionnalités riches permettent aux utilisateurs de tous niveaux de conduire des analyses statistiques professionnelles sans écrire une seule ligne de code. Les limites identifiées constituent des axes d'amélioration clairs qui, une fois implémentés, permettront à l'application de rivaliser avec des solutions commerciales établies. L'architecture modulaire actuelle facilite l'intégration progressive de ces nouvelles fonctionnalités. La roadmap proposée priorise les améliorations ayant le plus fort impact utilisateur : performance, enrichissement des analyses statistiques, et personnalisation des visualisations. Ces développements renforceront la position de EDA-Desk Pro comme outil de référence pour l'analyse exploratoire de données.