Portfolio d'Analyste de Données

Présentation Générale

Ce portfolio démontre mes compétences en analyse de données à travers quatre projets complets, couvrant les aspects essentiels du métier d'analyste de données junior. Chaque projet illustre une compétence technique différente et utilise des jeux de données réels provenant de sources reconnues comme Kaggle et GitHub.

Objectifs du Portfolio

- **Démontrer la maîtrise technique** : Utilisation de Python, pandas, matplotlib, seaborn, scikit-learn, NLTK, et statsmodels
- **Illustrer la méthodologie** : Application rigoureuse de méthodes d'analyse de données
- **Présenter la capacité de communication** : Visualisations claires et insights actionnables
- **Montrer la polyvalence** : Différents types d'analyses (exploratoire, prédictive, statistique, textuelle)

Compétences Techniques Démontrées

Langages et Outils

- **Python** : Langage principal pour toutes les analyses
- **Pandas** : Manipulation et nettoyage des données
- **Matplotlib & Seaborn** : Visualisation de données
- **Scikit-learn** : Machine learning et clustering
- **NLTK** : Traitement du langage naturel

- **Statsmodels** : Tests statistiques
- **React & Recharts** : Développement de tableaux de bord interactifs

Méthodologies

- Analyse exploratoire de données (EDA)
- Clustering et segmentation
- Classification de texte
- Tests d'hypothèses statistiques
- Visualisation de données
- Développement de tableaux de bord

Projet 1 : Analyse Exploratoire de Données (EDA)

Description

Analyse complète des performances des étudiants aux examens, explorant les facteurs qui influencent les résultats scolaires.

Jeu de Données

- **Source** : Kaggle - Students Performance in Exams
- **Taille** : 1000 observations
- **Variables** : Genre, groupe ethnique, niveau d'éducation parental, déjeuner, cours de préparation, scores en mathématiques, lecture et écriture

Méthodologie

1. **Exploration initiale** : Analyse des statistiques descriptives et identification des valeurs manquantes
2. **Analyse univariée** : Distribution de chaque variable
3. **Analyse bivariée** : Relations entre variables explicatives et scores
4. **Visualisations** : Histogrammes, boxplots, graphiques en barres

Insights Clés

- **Différences de genre** : Les filles obtiennent de meilleurs résultats en lecture (score moyen : 72.6) et écriture (score moyen : 72.5) que les garçons
- **Impact de l'éducation parentale** : Les étudiants dont les parents ont un niveau d'éducation supérieur obtiennent des scores significativement plus élevés
- **Efficacité de la préparation** : Les cours de préparation améliorent les scores de 5 à 10 points en moyenne dans toutes les matières

Compétences Démontrées

- Nettoyage et préparation des données
- Analyse statistique descriptive
- Création de visualisations informatives
- Interprétation des résultats et formulation d'insights

Projet 2 : Segmentation Client avec K-Means

Description

Application de l'algorithme K-Means pour segmenter la clientèle d'un centre commercial en fonction de leurs caractéristiques démographiques et comportementales.

Jeu de Données

- **Source** : Kaggle - Mall Customer Segmentation Data
- **Taille** : 200 clients
- **Variables** : ID client, genre, âge, revenu annuel, score de dépense

Méthodologie

1. **Préparation des données** : Nettoyage et normalisation

2. **Détermination du nombre optimal de clusters** : Méthode du coude (Elbow Method)
3. **Application de K-Means** : Clustering avec k=4
4. **Analyse des segments** : Caractérisation de chaque cluster
5. **Visualisation** : Graphiques de dispersion et diagrammes circulaires

Résultats

Identification de 4 segments clients distincts :

1. **Clients Cibles (35%)** : Revenus élevés (70k+), scores de dépense élevés (60+)
2. **Clients Conservateurs (25%)** : Revenus élevés, scores de dépense faibles
3. **Clients Économes (20%)** : Revenus faibles, scores de dépense faibles
4. **Clients Dépensiers (20%)** : Revenus faibles, scores de dépense élevés

Recommandations Business

- **Cibler les clients conservateurs** : Campagnes pour augmenter leur engagement
- **Fidéliser les clients cibles** : Programmes de fidélité premium
- **Développer des offres économiques** : Pour les segments à revenus faibles

Compétences Démontrées

- Algorithmes de machine learning non supervisé
- Optimisation des hyperparamètres
- Interprétation business des résultats techniques
- Formulation de recommandations stratégiques

Projet 3 : Analyse de Sentiments sur Données Textuelles

Description

Classification automatique de sentiments sur des critiques de films IMDB en utilisant des techniques de traitement du langage naturel.

Jeu de Données

- **Source** : Kaggle - IMDB Dataset of 50K Movie Reviews
- **Échantillon utilisé** : 5000 critiques (pour optimiser les performances)
- **Variables** : Texte de la critique, sentiment (positif/négatif)

Méthodologie

1. **Prétraitement du texte** :
2. Suppression des balises HTML
3. Conversion en minuscules
4. Suppression de la ponctuation
5. Élimination des mots vides (stopwords)
6. Stemming avec PorterStemmer
7. **Vectorisation** : TF-IDF (Term Frequency-Inverse Document Frequency)
8. **Modélisation** : Régression logistique pour la classification binaire
9. **Évaluation** : Matrice de confusion, précision, rappel, F1-score

Résultats

- **Précision du modèle** : 85%
- **Distribution équilibrée** : 50% sentiments positifs, 50% sentiments négatifs
- **Performance robuste** : Bonne généralisation sur les données de test

Défis Techniques

- **Gestion de la taille des données** : Optimisation pour traiter efficacement un large volume de texte
- **Prétraitement complexe** : Pipeline de nettoyage multi-étapes
- **Équilibrage des classes** : Maintien de la représentativité

Compétences Démontrées

- Traitement du langage naturel (NLP)
- Prétraitement de données textuelles
- Techniques de vectorisation
- Classification supervisée
- Évaluation de modèles de machine learning

Projet 4 : Test A/B et Analyse Statistique

Description

Analyse statistique rigoureuse d'un test A/B pour évaluer l'efficacité d'une nouvelle page d'atterrissage par rapport à l'ancienne version.

Jeu de Données

- **Source** : GitHub - A/B Test Data
- **Taille** : ~290,000 utilisateurs
- **Variables** : ID utilisateur, groupe (contrôle/traitement), page d'atterrissage, conversion

Méthodologie

1. **Nettoyage des données** :
2. Suppression des incohérences (page/groupe non alignés)
3. Élimination des doublons d'utilisateurs

4. Validation de l'intégrité des données
5. **Analyse descriptive** :
6. Calcul des taux de conversion par groupe
7. Vérification de l'équilibrage des groupes
8. **Test d'hypothèse** :
9. **H0** : Pas de différence significative entre les taux de conversion
10. **H1** : Différence significative entre les taux de conversion
11. Test Z pour la comparaison de proportions

Résultats

- **Groupe de contrôle** : 12.04% de taux de conversion
- **Groupe de traitement** : 11.88% de taux de conversion
- **Statistique Z** : 1.31
- **P-value** : 0.19 (> 0.05)

Conclusion Statistique

Avec une p-value de 0.19, nous ne pouvons pas rejeter l'hypothèse nulle. Il n'y a pas de preuve statistique suffisante pour affirmer que la nouvelle page améliore significativement le taux de conversion.

Recommandations

1. **Ne pas déployer** la nouvelle page dans son état actuel
2. **Itérer sur le design** pour identifier les éléments à améliorer
3. **Relancer un test** avec une version modifiée
4. **Considérer d'autres métriques** (temps sur page, taux de rebond)

Compétences Démontrées

- Tests d'hypothèses statistiques

- Analyse de la significativité
- Interprétation rigoureuse des résultats
- Formulation de recommandations basées sur les preuves
- Maîtrise des concepts de p-value et d'erreurs de type I/II

Tableau de Bord Interactif

Description

Développement d'une application web interactive présentant tous les projets du portfolio avec des visualisations dynamiques et une navigation intuitive.

Technologies Utilisées

- **Frontend** : React 18 avec Vite
- **Styling** : Tailwind CSS avec shadcn/ui
- **Visualisations** : Recharts (graphiques interactifs)
- **Icons** : Lucide React
- **Responsive Design** : Compatible desktop et mobile

Fonctionnalités

- **Navigation par onglets** : Accès rapide à chaque projet
- **Visualisations interactives** : Graphiques en barres, camemberts, histogrammes
- **Informations détaillées** : Compétences, insights, et méthodologies
- **Design professionnel** : Interface moderne et épurée
- **Statistiques globales** : Vue d'ensemble du portfolio

Architecture

```
src/
├── components/ui/     # Composants UI réutilisables
├── assets/           # Images et ressources
├── App.jsx          # Composant principal
└── App.css          # Styles personnalisés
```

Déploiement

L'application est construite pour la production avec optimisation automatique des performances et peut être déployée sur n'importe quelle plateforme d'hébergement statique.

Conclusion et Perspectives

Compétences Acquises

Ce portfolio démontre une maîtrise complète du cycle de vie de l'analyse de données :

1. **Collecte et préparation** : Acquisition de données depuis des sources fiables
2. **Exploration et nettoyage** : Identification et traitement des anomalies
3. **Analyse et modélisation** : Application de techniques statistiques et de machine learning
4. **Visualisation** : Création de graphiques informatifs et esthétiques
5. **Communication** : Présentation claire des résultats et recommandations
6. **Développement** : Création d'interfaces interactives pour la présentation

Valeur Ajoutée pour l'Entreprise

- **Prise de décision basée sur les données** : Capacité à transformer les données en insights actionnables
- **Polyvalence technique** : Maîtrise de multiples outils et techniques d'analyse
- **Communication efficace** : Aptitude à présenter des résultats complexes de manière accessible

- **Approche méthodologique** : Rigueur scientifique dans l'analyse et l'interprétation

Prochaines Étapes

- **Approfondissement du machine learning** : Exploration d'algorithmes plus avancés
- **Big Data** : Apprentissage d'outils comme Spark ou Hadoop
- **Cloud Computing** : Utilisation de plateformes comme AWS ou GCP
- **Automatisation** : Développement de pipelines de données automatisés

---

Ce portfolio a été créé pour démontrer mes compétences en analyse de données et ma capacité à mener des projets complets de bout en bout. Chaque projet illustre une facette différente du métier d'analyste de données et témoigne de ma passion pour transformer les données en valeur business.