LDA Advisory

12 Projets pour devenir Data Analyst

Natacha Njongwa Yepnga

LeCoinStat

Introduction

Ce notebook présente 12 projets concrets pour développer les compétences nécessaires en data analysis en utilisant Excel, Python, R, SQL, Power BI ou Tableau. Chaque projet est détaillé avec une problématique métier, des objectifs, les compétences à acquérir et les bases de données recommandées.

Projet 1 : Suivre et visualiser les performances des ventes avec un tableau de bord (Excel)

Problématique métier

Une entreprise souhaite suivre les performances des ventes par région et catégorie pour identifier les zones à améliorer.

Objectif

Créer un tableau de bord interactif pour analyser les KPI clés tels que les ventes, les marges, et les performances régionales.

Compétences à acquérir

- Nettoyage et organisation des données.
- Formules Excel essentielles : SOMME, SOMME.SI, MOYENNE, RECHERCHEV, INDEX, EQUIV.
- Création de tableaux croisés dynamiques avec des segments pour une navigation facile.
- Création de graphiques dynamiques et tableaux de bord interactifs.

Base de données

- Superstore Dataset

Projet 2 : Explorer et comprendre les performances des employés (Python ou R)

Problématique métier

Une entreprise souhaite analyser la répartition des performances des employés pour comprendre les écarts et identifier les outliers.

Objectif

Étudier les distributions des scores de performance et des heures travaillées pour détecter les facteurs d'amélioration.

Compétences à acquérir

- Statistiques descriptives : Moyenne, médiane, mode, quartiles, variance, écart-type.
- Visualisation des distributions : Histogrammes, boxplots, pie chart, diagramme en barre.
- Détection et analyse des outliers avec la règle des 1.5 * IQR.
- Manipulation des données avec Python (`pandas`) ou R (`dplyr`).

Base de données

- HR Analytics Dataset

Projet 3 : Identifier les facteurs qui influencent les performances commerciales (Python ou R)

Problématique métier

Une entreprise souhaite comprendre les interactions entre ses produits, ses performances régionales et ses marges pour optimiser sa stratégie commerciale.

Objectif

Analyser les relations quali-quali, quali-quanti et quanti-quanti pour détecter des facteurs clés.

Compétences à acquérir

- Analyse bivariée descriptive : Corrélations, tableaux croisés, comparaisons de moyennes.
- Visualisation avancée : Scatter plots, boxplots, heatmap.
- Interprétation des relations pour guider les décisions stratégiques.

Base de données

- Superstore Dataset

Ces projets vous inspirent ? Suivez-moi sur mes différents réseaux pour ne pas rater les prochains.

Projet 4 : Analyser les écarts de performance entre départements (Python ou R)

Problématique métier

Une entreprise souhaite analyser les écarts de performance entre ses départements en explorant les relations entre variables qualitatives (départements, satisfaction) et quantitatives (scores de performance, heures travaillées).

Objectif

Valider les différences observées en appliquant des tests statistiques adaptés.

Compétences à acquérir

Tests statistiques à maîtriser

1. Tests pour les relations qualitatives (quali-quali)

- **Khi-deux (Chi-Square Test)** : Vérifie l'indépendance entre deux variables qualitatives.

  ***Exemple** : La satisfaction des employés dépend-elle du département ?*

- **V de Cramer :** Mesure la force de l'association entre deux variables qualitatives. À utiliser après un test Khi-deux.

  ***Exemple** : Quelle est la force de l'association entre le département et la satisfaction ?*
  ***Interprétation** : Les valeurs vont de 0 (aucune association) à 1 (association parfaite)*

- **T de Tschuprow** : Alternative au V de Cramer, adaptée pour des tableaux asymétriques (nombre de lignes et colonnes déséquilibré).

- **Fisher Exact Test :** Alternative au Khi-deux pour des échantillons de petite taille.

2. Tests pour comparer les moyennes (quali-quanti)

- **ANOVA (Analyse de Variance) :** Compare les moyennes entre plus de deux groupes (données normalement distribuées et variances homogènes).

  ***Exemple** : Les scores de performance diffèrent-ils selon le département ?*

- **Kruskal-Wallis :** Alternative non paramétrique à l'ANOVA pour des données non normalement distribuées.

- **Test t de Student :** Compare les moyennes entre deux groupes (données normalement distribuées).

- **Wilcoxon Rank-Sum (ou Mann-Whitney U)** : Alternative non paramétrique au test t pour deux groupes.

3. Tests de variances (homogénéité des variances)

- **Levene Test** : Vérifie si les variances entre groupes sont homogènes.
- **Bartlett's Test** : Vérifie l'homogénéité des variances (utilisé avant l'ANOVA).

4. Tests de normalité

- **Shapiro-Wilk :** Vérifie si un échantillon suit une distribution normale.
- **Kolmogorov-Smirnov** : Vérifie si un échantillon suit une distribution spécifique (normale ou autre).
- **Jarque-Bera** : Teste la normalité en fonction de la symétrie (skewness) et de l'aplatissement (kurtosis).

5. Tests pour les relations quantitatives (quanti-quanti)

- **Corrélation de Pearson** : Mesure la relation linéaire entre deux variables continues (si elles sont normalement distribuées).
- **Corrélation de Spearman :** Alternative non paramétrique à Pearson, adaptée pour des relations monotones.
- **Corrélation de Kendall Tau :** Mesure la relation monotone entre deux variables continues ou ordinales.

6. Tests non paramétriques complémentaires

- **Test de Wilcoxon signé-rang :** Compare des paires pour des données non paramétriques (équivalent non paramétrique du t-test apparié).
- **Friedman Test :** Alternative non paramétrique pour l'ANOVA à mesures répétées.

Visualisation des résultats : Boxplots, heatmaps, scatter plots.

Interprétation des résultats et formulation de recommandations métier.

Base de données

- Employee Performance Evaluation Dataset

Ces projets vous inspirent ? Suivez-moi sur mes différents réseaux pour ne pas rater les prochains.

Projet 5 : Automatiser les rapports de ventes avec SQL

Problématique métier

Une entreprise souhaite produire un rapport détaillant les ventes totales par produit pour le dernier trimestre.

Objectif

Utiliser SQL pour extraire, filtrer et agréger les données nécessaires.

Compétences à acquérir

- Requêtes SQL de base : SELECT, WHERE, GROUP BY, HAVING.
- Calculs d'agrégats : SOMME, MOYENNE pour des KPI.
- Automatisation des rapports avec des vues SQL.

Base de données

- Chinook Database

Exercices SQL avec la base de données Chinook

Introduction

Consolidons vos compétences en requêtes SQL avec un ensemble d'exercices qui mettront vos connaissances à l'épreuve. Gardez la base de données **Chinook** et l'outil **DB Browser for SQLite** à portée de main.

Pour chaque exercice, fournissez la requête SQL appropriée et conservez vos réponses dans un fichier nommé chinook-queries.sql.

Exigences

- Utilisez la base de données Chinook.
- Travaillez avec DB Browser for SQLite ou tout autre environnement de votre choix.

Exercices SQL

Requêtes de base

1. **Clients non américains** : Fournissez une requête affichant les Clients (leurs noms complets, ID client et pays) qui ne sont pas aux États-Unis.

2. **Clients brésiliens :** Fournissez une requête affichant uniquement les Clients provenant du Brésil.

3. **Factures des clients brésiliens** : Fournissez une requête affichant les factures des clients qui sont du Brésil.

Le tableau résultant doit inclure le nom complet du client, l'ID de la facture, la date de la facture et le pays de facturation.

4. **Agents de vente :** Fournissez une requête affichant uniquement les employés qui sont des Agents de Vente

Agrégations et relations

5. **Pays uniques dans les factures** : Fournissez une requête affichant une liste unique des pays de facturation présents dans la table Invoice.

6. **Factures par agent de vente :** Fournissez une requête affichant les factures associées à chaque agent de vente.

   Le tableau résultant doit inclure le nom complet de l'agent de vente.

7. **Détails des factures** : Fournissez une requête affichant le total de chaque facture, le nom du client, le pays et le nom de l'agent de vente.

Analyse par année et lignes de facture

8. Ventes par année : Combien de factures y a-t-il eu en 2009 et 2011 ? Quels sont les montants totaux des ventes pour chacune de ces années ?

9. Articles pour une facture donnée : Fournissez une requête comptant le nombre d'articles (line items) pour l'ID de facture 37.

10. Articles par facture : Fournissez une requête comptant le nombre d'articles (line items) pour chaque facture.

Astuce : utilisez GROUP BY.

Détails des morceaux

11. Nom des morceaux : Fournissez une requête incluant le nom du morceau pour chaque ligne de facture.

12. Morceaux et artistes : Fournissez une requête incluant le nom du morceau acheté ET le nom de l'artiste pour chaque ligne de facture.

Comptages et regroupements

13. Nombre de factures par pays : Fournissez une requête affichant le nombre de factures par pays.

Astuce : utilisez GROUP BY.

14. Nombre de morceaux par playlist : Fournissez une requête affichant le nombre total de morceaux dans chaque playlist.
Le nom de la playlist doit être inclus dans le tableau résultant.

15. Liste des morceaux : Fournissez une requête affichant tous les morceaux (Tracks), mais sans afficher les IDs.
Le tableau résultant doit inclure le nom de l'album, le type de média et le genre.

Analyse des ventes

16. Factures et articles : Fournissez une requête affichant toutes les factures, avec le nombre d'articles par facture.

17. Ventes par agent de vente : Fournissez une requête affichant les ventes totales réalisées par chaque agent de vente.

18. Meilleur agent de 2009 : Quel agent de vente a réalisé le plus de ventes en 2009 ?

19. Meilleur agent de 2010 : Quel agent de vente a réalisé le plus de ventes en 2010 ?

20. Meilleur agent global : Quel agent de vente a réalisé le plus de ventes en tout ?

Analyse des clients et des pays

21. Clients par agent de vente : Fournissez une requête affichant le nombre de clients attribués à chaque agent de vente.

22. Ventes totales par pays : Fournissez une requête affichant les ventes totales par pays.
Quel pays a dépensé le plus ?

Analyse des morceaux et des artistes

23. Morceau le plus acheté en 2013 : Fournissez une requête affichant le morceau le plus acheté en 2013.

24. Top 5 des morceaux les plus achetés : Fournissez une requête affichant les 5 morceaux les plus achetés en tout.

25. Top 3 des artistes les plus vendus : Fournissez une requête affichant les 3 artistes les plus vendus.

26. Type de média le plus acheté : Fournissez une requête affichant le type de média le plus acheté.

Ressources supplémentaires

- SQL Course
- Cheatsheet SQL sur GitHub
- SQL Cheatsheet
- Sololearn - Cours SQL

Source des exercices

Ces exercices sont adaptés de la ressource suivante :

GitHub - LucasMcL/15-sql_queries_02-chinook

Ces projets vous inspirent ? Suivez-moi sur mes différents réseaux pour ne pas rater les prochains.

Projet 6 Analyser le profil des vins : Réduction de la dimensionnalité avec l'ACP

Problématique métier

Une entreprise souhaite identifier les axes clés de différenciation entre ses clients afin de mieux personnaliser ses offres. Ce projet explore un ensemble de données sur la qualité du vin pour identifier ces différences.

Compétences à acquérir

- Réduction de dimensionnalité : Utilisation de l'ACP pour les données continues .
- Visualisation : Réduction des dimensions pour visualiser les données en 2D et 3D.
- Interprétation des résultats : Comprendre les axes discriminants et les groupes significatifs.

Base de données

Le dataset utilisé dans ce projet est le Wine Quality Dataset disponible sur UCI Machine Learning Repository. Ce dataset contient des informations sur les propriétés chimiques de différents vins ainsi que leur qualité, notée de 0 à 10. Le dataset est accessible via le lien suivant :

Wine Quality Dataset - UCI Machine Learning Repository

Compétences techniques nécessaires

- Analyse en Composantes Principales (ACP)
- Visualisation des données (matplotlib, seaborn)
- Traitement de données avec pandas et numpy

Projet 7 : Créer un tableau de bord interactif pour les performances commerciales (Power BI ou Tableau)

Problématique métier

Une entreprise souhaite un tableau de bord interactif pour suivre les performances commerciales en temps réel.

Objectifs

Visualiser les KPI clés, les tendances de ventes et les performances régionales.

Compétences à acquérir

- Création de tableaux de bord interactifs avec slicers et filtres dynamiques.
- Visualisation avancée : Graphiques temporels, cartographiques, combinés.
- Synthèse et présentation des données pour une prise de décision rapide.

Base de données

Global Superstore Dataset

Projet 8 SQL Avancé : Analyse des performances de l'entreprise avec la base de données AdventureWorks

Problématique métier

Une entreprise souhaite analyser ses performances en termes de ventes, de produits, de régions et de segments de clients. L'objectif est de produire des rapports automatisés et de segmenter les clients et les ventes afin d'optimiser les décisions commerciales.

Objectif

Utiliser SQL pour :

- Analyser les ventes par produit, région et segment de client.
- Identifier les produits les plus rentables et les moins performants.
- Calculer des indicateurs clés de performance (KPI) comme le revenu moyen par client, le revenu total par agent de vente, etc.
- Créer des vues SQL pour automatiser la génération de rapports de performance.

Compétences à acquérir

- Jointures complexes (INNER JOIN, LEFT JOIN, RIGHT JOIN).
- Utilisation des fenêtres (Window Functions) pour des calculs cumulés et de moyennes mobiles.
- Création de sous-requêtes et de vues SQL pour automatiser la génération de rapports.
- Agrégations de données : SUM, AVG, COUNT, etc.
- Utilisation de GROUP BY pour l'agrégation des données par différentes dimensions.

Base de données

Vous utiliserez la base de données AdventureWorks, disponible sur GitHub - AdventureWorks Database.

Exercices SQL

1. Requêtes de base

1.Ventes par région : Afficher le total des ventes par région pour l'année 2024.

2.Clients actifs : Fournir une requête affichant les clients actifs (ceux ayant passé des commandes récemment).

3.Ventes par catégorie de produit : Afficher les ventes totales pour chaque catégorie de produit.

2. Agrégations et relations

4.Ventes par agent de vente : Afficher les ventes totales réalisées par chaque agent de vente.

5.Top 5 des produits : Afficher les 5 produits les plus rentables (en termes de revenus générés).

6.Ventes par segment de client : Fournir une requête affichant le total des ventes pour chaque segment de clients.

3. Analyse des segments de clients

7.Segmentation des clients par montant dépensé : Créer des segments de clients basés sur le montant total dépensé (par exemple, bas, moyen, élevé).

8.Top 10 des clients : Identifier les 10 clients ayant généré le plus de revenus.

9.Clients par type de commande : Fournir une requête affichant les clients qui ont acheté des produits en ligne et ceux qui ont fait des achats en magasin.

4. Analyse par produit

10.Produits les plus populaires : Identifier les produits les plus achetés au cours des 12 derniers mois.

11.Ventes par type de produit : Fournir une requête affichant les ventes par type de produit (par exemple, électronique, vêtements, etc.).

12. Ventes totales par produit et par mois : Afficher les ventes totales de chaque produit sur une période de 12 mois.

5. Optimisation des ventes

13.Ventes par région et par produit : Afficher les ventes totales par produit pour chaque région.

14.Taux de retour des produits : Calculer le taux de retour des produits (nombre d'articles retournés divisé par le nombre d'articles vendus).

15.Rentabilité par produit : Identifier les produits les plus rentables et ceux qui ne génèrent pas suffisamment de revenus.

Ressources supplémentaires

- AdventureWorks Database Documentation
- SQL Course
- SQL Cheat Sheet

Source des exercices

Ce projet SQL avancé vous permet d'analyser les performances de l'entreprise à l’aide de la base de données AdventureWorks, une base de données représentant une entreprise de vente au détail. Vous y apprendrez à manipuler les données pour obtenir des informations précieuses sur les ventes, les clients, les produits et la rentabilité. Vous pourrez automatiser la génération de rapports, effectuer des analyses de performance, et segmenter les données de manière pertinente.

Ces projets vous inspirent ? Suivez-moi sur mes différents réseaux pour ne pas rater les prochains.

Projet 9 : Segmenter les clients avec des algorithmes de clustering (Python ou R)

Problématique métier

Une entreprise souhaite regrouper ses clients en segments homogènes pour personnaliser ses offres.

Objectifs

Utiliser des algorithmes de clustering pour identifier des groupes distincts.

Compétences à acquérir

- Préparation des données pour le clustering : Normalisation, nettoyage.
- Implémentation de K-means,CAH, DBSCAN ou autres algorithmes.
- Visualisation des résultats en 2D ou 3D.

Base de données

Mall Customer Segmentation Dataset

Projet 10 : Prédire les prix des maisons avec une régression linéaire (Python ou R)

Problématique métier

Une entreprise souhaite prédire les prix des maisons en fonction de leurs caractéristiques.

Objectifs

Construire un modèle de régression linéaire pour analyser les facteurs influents.

Compétences à acquérir

- Modélisation prédictive avec régression linéaire simple et multiple.
- Évaluation des performances : R^2,RMSE, MSE, MAPE, analyse des résidus.
- Interprétation des coefficients pour détecter les facteurs clés.

Base de données

Boston Housing Dataset

Projet 11 : Prédire les défauts de paiement avec une régression logistique (Python ou R)

Problématique métier

Une entreprise souhaite prédire si un client fera défaut sur son paiement en fonction de ses caractéristiques.

Objectif

Construire un modèle de classification binaire pour identifier les clients à risque.

Compétences à acquérir

- Classification binaire avec régression logistique.
- Évaluation des performances : Matrice de confusion, AUC, F1-score.
- Interprétation des probabilités pour la prise de décision.

Base de données

Credit Card Default Dataset

Projet 12 : Prévoir les ventes avec des séries temporelles (Python ou R)

Problématique métier

Une entreprise souhaite prévoir les ventes pour ajuster ses stocks et ressources.

Objectif

Utiliser une série temporelle pour analyser les tendances et prédire les performances futures.

Compétences à acquérir

- Analyse exploratoire des séries temporelles : Identification des tendances, saisonnalité et bruit.
- Modélisation avec ARIMA ou Prophet.
- Visualisation des prévisions pour guider les décisions stratégiques.

Base de données

Store Sales - Time Series Forecasting

Ces projets vous inspirent ? Suivez-moi sur mes différents réseaux pour ne pas rater les prochains.

Pour ne pas manquer mes prochains posts, suivez moi sur LinkedIn, YouTube, Instagram et TikTok