Test Global Data Analyst - Niveau Intermédiaire (2/4)

Contexte: Entretien pour Data Analyst - UniBank Haiti
Durée estimée: 45-60 minutes
Nombre de questions: 35

Section A: EDA et Data Wrangling (10 questions)

Q1. Décrivez votre processus d'Exploratory Data Analysis (EDA).

R1. 1. **Comprendre le contexte business** - Objectifs, stakeholders 2. **Charger et examiner la structure** - shape, dtypes, head() 3. **Profiler chaque variable** - describe(), distributions 4. **Identifier les problèmes de qualité** - manquants, outliers, doublons 5. **Explorer les relations** - corrélations, cross-tabulations 6. **Documenter les insights** - observations, recommandations

---

Q2. Comment traiteriez-vous les valeurs manquantes dans un dataset de clients bancaires?

R2. Dépend du contexte: 1. **Identifier le type:** MCAR (aléatoire), MAR (dépendant d'autres variables), MNAR (dépendant de la valeur elle-même) 2. **Évaluer la proportion:** $< 5\% \rightarrow$ suppression possible; $> 50\% \rightarrow$ supprimer la colonne 3. **Stratégies:** - Imputation par médiane (numérique, robuste) - Imputation par mode (catégoriel) - Imputation par groupe - Suppression si peu nombreux et aléatoires

---

Q3. Quelle est la différence entre MCAR, MAR et MNAR?

R3. - **MCAR (Missing Completely At Random):** Manquant indépendant de toutes les variables - **MAR (Missing At Random):** Manquant dépend d'autres variables observées - **MNAR (Missing Not At Random):** Manquant dépend de la valeur elle-même

Exemple: Revenu manquant car haut revenus ne veulent pas déclarer = MNAR

Q4. Comment détecteriez-vous les doublons dans un dataset?

R4.

```
# Doublons complets
df.duplicated().sum()
df[df.duplicated()]

# Doublons sur colonnes clés
df[df.duplicated(subset=['client_id', 'date'])]

# Avant de supprimer, toujours investiguer!
# Parfois les doublons sont légitimes
```

Q5. Qu'est-ce que le feature engineering? Donnez des exemples bancaires.

R5. Feature engineering = Création de nouvelles variables à partir des données existantes.

Exemples bancaires: - **Ratio dette/revenu (DTI):** dette_totale / revenu - **Ancienneté:** (aujourd'hui - date_inscription).days - **Taux d'utilisation crédit:** solde_utilisé / limite -

RFM scores: Recency, Frequency, Monetary - **Indicateurs binaires:** is_nouveau_client, has_defaut_passé

Q6. Comment normaliseriez-vous des données et pourquoi?

R6. Pourquoi: Mettre les variables sur une même échelle pour comparaison ou pour certains algorithmes.

Méthodes: - **Min-Max scaling:** (x - min) / (max - min) → [0, 1] - **Z-score standardisation:** (x - μ) / σ → moyenne 0, écart-type 1 - **Log transformation:** Pour données asymétriques positives

---

Q7. Comment valideriez-vous la qualité de vos données?

R7. 1. **Complétude:** Vérifier les valeurs manquantes 2. **Unicité:** Vérifier les doublons 3. **Validité:** Valeurs dans les plages attendues 4. **Cohérence:** Cross-validation entre variables 5. **Exactitude:** Comparaison avec sources de référence

```
# Exemples d'assertions
assert df['montant'].min() >= 0, "Montants négatifs"
assert df['taux'].between(0, 1).all(), "Taux hors limites"
assert (df['date_fin'] >= df['date_debut']).all(), "Dates incohérentes"
```

Q8. Comment gérez-vous les outliers dans un contexte bancaire?

R8. 1. **Identifier:** Box plot, IQR method, Z-score 2. **Investiguer:** Erreur ou valeur réelle? 3. **Décider:** - **Garder:** Si valeur légitime (gros client) - **Corriger:** Si erreur de saisie - **Capper:** Limiter à un seuil - **Exclure:** Analyse séparée

Important en banque: Les outliers peuvent être des clients VIP ou des fraudes - ne jamais supprimer aveuglément.

Q9. Qu'est-ce qu'un tableau de contingence et quand l'utiliser?

R9. Tableau de contingence = tableau croisé de fréquences entre deux variables catégorielles.

Usage: - Analyser la relation entre deux variables catégorielles - Base pour le test Chi-carré - Visualiser les proportions croisées

```
pd.crosstab(df['segment'], df['produit'], normalize='index')
```

Q10. Comment interprétez-vous une matrice de corrélation?

R10. - **Diagonale = 1:** Variable corrélée avec elle-même - **Valeurs proches de 1 ou -1:** Forte corrélation (potentielle multicolinéarité) - **Valeurs proches de 0:** Pas de relation linéaire

Actions: - Identifier les variables très corrélées (>0.7 ou <-0.7) - Considérer supprimer ou combiner les variables redondantes - Investiguer les corrélations inattendues

---

Section B: Visualisation de Données (8 questions)

Q11. Quel graphique utiliseriez-vous pour comparer les distributions de soldes entre segments de clients?

R11. **Box plot** ou **violin plot** par segment. - Box plot: Montre médiane, quartiles, outliers - Violin plot: Ajoute la forme de la distribution

Alternative: Histogrammes superposés avec transparence.

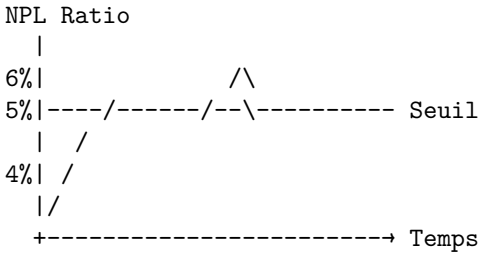
Q12. Pourquoi un pie chart n'est-il pas toujours approprié?

R12. - Difficile de comparer des parts similaires - Limité à peu de catégories (5-6 max) - Pas adapté pour montrer l'évolution - Ne permet pas les valeurs négatives

Alternative: Bar chart horizontal, souvent plus lisible.

Q13. Comment visualiseriez-vous l'évolution du NPL ratio sur 3 ans?

R13. **Line chart** avec: - Axe X: Temps (mois/trimestre) - Axe Y: NPL ratio (%) - Ligne de seuil à 5% (benchmark) - Optionnel: Bandes de confiance ou zone de danger



Q14. Quand utiliseriez-vous un scatter plot vs un heatmap?

R14. - **Scatter plot:** Visualiser la relation entre 2 variables continues (peu de points à milliers) - **Heatmap:** - Matrice de corrélation (nombreuses variables) - Données avec 3 dimensions (ex: heure × jour × transactions) - Trop de points pour un scatter (utiliser hexbin)

Q15. Comment visualiseriez-vous la composition du portefeuille de prêts par secteur et par niveau de risque?

R15. Options: 1. **Stacked bar chart:** Secteurs en X, couleurs par niveau de risque 2. **Treemap:** Taille = montant, couleur = risque 3. **Heatmap:** Secteur × Risque, intensité = montant

Le treemap est particulièrement efficace pour les données hiérarchiques.

Q16. Quelles sont les bonnes pratiques pour un dashboard exécutif?

R16. 1. **KPIs en haut:** Indicateurs clés avec tendance 2. **5-7 visualisations max:** Éviter la surcharge 3. **Hiérarchie visuelle:** Important = grand et en haut 4. **Cohérence:** Mêmes couleurs pour mêmes concepts 5. **Contexte:** Benchmarks, périodes de comparaison 6. **Interactivité:** Filtres pour drill-down

---

Q17. Comment assurer l'accessibilité de vos visualisations?

R17. - Éviter rouge/vert seuls (daltonisme) - Utiliser des patterns en plus des couleurs - Étiquettes et légendes claires - Taille de police suffisante - Alt-text pour les rapports digitaux - Contraste suffisant

---

Q18. Qu'est-ce qu'un sparkline et quand l'utiliser?

R18. Sparkline = mini-graphique intégré dans une cellule ou un tableau.

Usage: - Dashboard avec beaucoup de lignes - Montrer la tendance sans détails - Tableaux comparatifs

Exemple: Liste des agences avec sparkline de performance mensuelle.

Section C: Probabilités et Statistiques Inférentielles (8 questions)

Q19. Expliquez le théorème de Bayes avec un exemple bancaire.

R19. Bayes: $P(A|B) = P(B|A) \times P(A) / P(B)$

Exemple fraude: - P(Fraude) = 1% - P(Alerte|Fraude) = 95% - P(Alerte|Non-fraude) = 5%

$P(Fraude|Alerte) = (0.95 \times 0.01) / (0.95 \times 0.01 + 0.05 \times 0.99) = 16\%$

Même avec une bonne détection, seulement 16% des alertes sont de vraies fraudes.

Q20. Qu'est-ce que le Théorème Central Limite et pourquoi est-il important?

R20. Le TCL dit que pour $n \geq 30$, la distribution des moyennes d'échantillons tend vers une normale, quelle que soit la distribution originale.

Importance: - Permet d'utiliser la normale pour l'inférence - Justifie les intervalles de confiance - Permet les tests d'hypothèses

---

Q21. Comment calculeriez-vous la taille d'échantillon pour une enquête de satisfaction?

R21.

$n = (z^2 \times p \times (1-p)) / m^2$

```
z = 1.96 pour 95% de confiance
p = proportion estimée (0.5 si inconnue)
m = marge d'erreur souhaitée

Exemple: IC 95%, marge 3%, p=0.5
n = (1.96² × 0.5 × 0.5) / 0.03² = 1068 clients
```

Q22. Quelle distribution utiliseriez-vous pour modéliser le nombre de fraudes par jour?

R22. Distribution de Poisson car: - Événements rares - Sur une période donnée - Indépendants les uns des autres

Paramètre λ = nombre moyen de fraudes par jour $P(X=k) = (\lambda^k \times e^{-\lambda}) / k!$

Q23. Comment testeriez-vous si deux segments ont des taux de défaut différents?

R23. Test z de deux proportions: - H_0: $p_1 = p_2$ - H_1: $p_1 \neq p_2$

Ou **Test Chi-carré** sur le tableau de contingence: | | Défaut | Non-défaut | |–|——–|———–| | Segment A | n_{11} | n_{12} | | Segment B | n_{21} | n_{22} |

Q24. Qu'est-ce que la puissance statistique d'un test?

R24. Puissance = 1 - β = Probabilité de détecter un effet s'il existe réellement.

Facteurs augmentant la puissance: - Taille d'échantillon plus grande - Taille de l'effet plus grande - Niveau α plus élevé - Variance plus faible

Objectif typique: 80% de puissance.

Q25. Comment interprétez-vous un coefficient de corrélation de -0.65?

R25. - **Signe négatif:** Relation inverse (quand X augmente, Y diminue) - **Magnitude 0.65:** Corrélation modérée à substantielle - **Interprétation:** Environ 42% de la variance partagée ($r^2 = 0.42$)

Attention: Corrélation ≠ Causalité. Il peut y avoir une variable confondante.

Q26. Qu'est-ce que l'Expected Loss (EL) en risque de crédit?

R26.

```
EL = PD × LGD × EAD

PD: Probability of Default (probabilité de défaut)
LGD: Loss Given Default (% de perte si défaut)
EAD: Exposure at Default (montant exposé)

Exemple:
PD = 5%, LGD = 45%, EAD = 100,000 HTG
EL = 0.05 × 0.45 × 100,000 = 2,250 HTG
```

Section D: Power BI / Python (9 questions)

Q27. Écrivez une mesure DAX pour calculer une moyenne mobile sur 3 mois.

R27.

```
MM3 = AVERAGEX(
    DATESINPERIOD(
        Calendrier[Date],
        MAX(Calendrier[Date]),
        -3,
        MONTH
    ),
    CALCULATE(SUM(Ventes[Montant]))
)
```

Q28. Comment créer un ranking en DAX?

R28.

```
Rang = RANKX(
    ALL(Clients),       -- Table de classement
    [Total Ventes],     -- Expression à classer
    ,                   -- Valeur alternative (vide)
    DESC,               -- Ordre décroissant
    Dense               -- Dense = ne saute pas les rangs
)
```

Q29. Comment filtrer un DataFrame Pandas pour obtenir les clients avec un solde > 100K et un segment “Premium”?

R29.

```python
# Méthode 1: Conditions
df_filtered = df[(df['solde'] > 100000) & (df['segment'] == 'Premium')]

# Méthode 2: query
df_filtered = df.query('solde > 100000 and segment == "Premium"')
```

Q30. Comment calculer des statistiques groupées en Pandas?

R30.

```python
# Agrégation simple
df.groupby('agence')['montant'].agg(['sum', 'mean', 'count'])

# Agrégation multiple
df.groupby('agence').agg({
    'montant': ['sum', 'mean'],
    'client_id': 'nunique',
    'date': ['min', 'max']
})
```

Q31. Comment pivoteriez-vous un DataFrame pour avoir les agences en lignes et les mois en colonnes?

R31.

```python
pivot = df.pivot_table(
    values='montant',
    index='agence',
    columns='mois',
    aggfunc='sum',
    fill_value=0
)
```

Q32. Comment créer une colonne calculée en Pandas basée sur plusieurs conditions?

R32.

```python
# Méthode 1: np.select
conditions = [
    df['score'] >= 700,
    df['score'] >= 500,
```

```
    df['score'] >= 300
]
choices = ['Low Risk', 'Medium Risk', 'High Risk']
df['risk_category'] = np.select(conditions, choices, default='Very High Risk')

# Méthode 2: apply
df['risk_category'] = df['score'].apply(
    lambda x: 'Low' if x >= 700 else ('Medium' if x >= 500 else 'High')
)
```

Q33. Comment joindre deux DataFrames en Pandas?

R33.

```
# Inner join (par défaut)
df_merged = pd.merge(df1, df2, on='client_id')

# Left join
df_merged = pd.merge(df1, df2, on='client_id', how='left')

# Avec colonnes différentes
df_merged = pd.merge(df1, df2, left_on='id_client', right_on='client_id')

# Plusieurs clés
df_merged = pd.merge(df1, df2, on=['client_id', 'date'])
```

Q34. Comment gérer les dates en Pandas?

R34.

```
# Conversion
df['date'] = pd.to_datetime(df['date'])

# Extraction
df['year'] = df['date'].dt.year
df['month'] = df['date'].dt.month
df['day_of_week'] = df['date'].dt.dayofweek

# Différence
df['ancienneté'] = (pd.Timestamp.now() - df['date']).dt.days

# Resampling
monthly = df.set_index('date').resample('M')['montant'].sum()
```

Q35. Expliquez l'intérêt d'utiliser Python dans Power BI.

R35. Python dans Power BI permet: 1. **Source de données:** Charger et transformer avec Pandas 2. **Visuels personnalisés:** Matplotlib, Seaborn, Plotly 3. **Analyses avancées:** Stats, ML, non disponibles en DAX 4. **Reproductibilité:** Code réutilisable

Cas d'usage: Heatmap de corrélation, analyses statistiques complexes, clustering.

Scoring

Score	Niveau
0-12	À améliorer
13-20	Débutant
21-28	Intermédiaire
29-33	Avancé
34-35	Expert