Test Global Data Analyst - Niveau Intermédiaire (1/4)

Contexte: Entretien pour Data Analyst - UniBank Haiti
Durée estimée: 45-60 minutes
Nombre de questions: 40

Section A: Statistiques Descriptives (10 questions)

Q1. Quelle mesure de tendance centrale utiliseriez-vous pour analyser les revenus des clients sachant qu'il y a quelques clients très riches?

R1. La **médiane**, car elle est robuste aux valeurs extrêmes (outliers). Les revenus sont typiquement asymétriques avec une queue à droite, ce qui tire la moyenne vers le haut. La médiane donne une meilleure représentation du client "typique".

Q2. La distribution des montants de transactions a un skewness de +1.8. Qu'est-ce que cela signifie?

R2. La distribution est **asymétrique positive** (queue à droite). Cela indique: - La majorité des transactions sont de petits montants - Quelques transactions de gros montants tirent la moyenne - Moyenne > Médiane > Mode - Un skewness > 1 est considéré comme fortement asymétrique

---

Q3. Comment détecteriez-vous les outliers dans un dataset de soldes de comptes?

R3. Deux méthodes principales: 1. **Méthode IQR:** Outlier si valeur < Q1 - 1.5×IQR ou > Q3 + 1.5×IQR 2. **Méthode Z-score:** Outlier si |Z| > 3 (plus de 3 écarts-types de la moyenne)

La méthode IQR est préférée car elle est robuste aux outliers eux-mêmes.

Q4. Quelle est la différence entre variance et écart-type?

R4. - **Variance (s^2):** Moyenne des carrés des écarts à la moyenne. Unité = carré de l'unité originale. - **Écart-type (s):** Racine carrée de la variance. Même unité que les données originales.

L'écart-type est plus interprétable. Exemple: Si les montants sont en HTG, la variance est en HTG^2, l'écart-type est en HTG.

Q5. Qu'est-ce que le coefficient de variation (CV) et quand l'utiliser?

R5. CV = (Écart-type / Moyenne) × 100%

Utilité: Comparer la dispersion relative de distributions avec des moyennes différentes.

Exemple: Agence A (moyenne 50K, σ=10K, CV=20%) vs Agence B (moyenne 200K, σ=30K, CV=15%). Malgré un écart-type plus élevé, l'agence B est plus homogène.

Q6. Que représentent les quartiles Q1, Q2, Q3?

R6. - **Q1 (25e percentile):** 25% des données sont inférieures - **Q2 (50e percentile):** La médiane, 50% en dessous - **Q3 (75e percentile):** 75% des données sont inférieures

L'IQR (Q3-Q1) contient 50% des données centrales.

Q7. Si la moyenne d'un échantillon est 100 et l'écart-type est 15, que peut-on dire des valeurs entre 70 et 130 si la distribution est normale?

R7. Ces valeurs correspondent à $\mu \pm 2\sigma$ (100 ± 30). Selon la règle empirique 68-95-99.7, environ **95%** des données se trouvent dans cet intervalle.

Q8. Qu'est-ce que le kurtosis et comment l'interpréter?

R8. Le kurtosis mesure l'épaisseur des queues de la distribution: - **Kurtosis = 3 (ou excess = 0):** Mésokurtique, comme la normale - **Kurtosis > 3:** Leptokurtique, queues épaisses, pic pointu - **Kurtosis < 3:** Platykurtique, queues fines, pic aplati

Un kurtosis élevé en finance indique plus d'événements extrêmes que prévu.

Q9. Comment résumez-vous une variable catégorielle?

R9. - **Mode:** Catégorie la plus fréquente - **Fréquences absolues:** Comptage par catégorie - **Fréquences relatives:** Pourcentage par catégorie - **Tableau de fréquences:** Distribution complète - **Visualisation:** Bar chart ou pie chart (si peu de catégories)

Q10. Quelle est la différence entre une variable discrète et continue?

R10. - **Discrète:** Valeurs dénombrables, entières (nb transactions, nb produits) - **Continue:** Valeurs sur un intervalle, mesurables (montant, taux, temps)

Impact: Les méthodes statistiques diffèrent (ex: histogramme pour continue, bar chart pour discrète).

Section B: Tests d'Hypothèses (8 questions)

Q11. Qu'est-ce qu'une p-value et comment l'interpréter?

R11. La p-value est la probabilité d'obtenir un résultat aussi extrême que celui observé, si l'hypothèse nulle (H_0) est vraie.

Interprétation: - $p < \alpha$ (souvent 0.05): Rejeter H_0, résultat significatif - $p \geq \alpha$: Ne pas rejeter H_0

Attention: p-value faible ≠ effet important. La significativité statistique ≠ significativité pratique.

Q12. Quelle est la différence entre erreur de Type I et Type II?

R12. - **Type I (α):** Faux positif - Rejeter H_0 alors qu'elle est vraie - **Type II (β):** Faux négatif - Ne pas rejeter H_0 alors qu'elle est fausse

La puissance = $1 - \beta$ = probabilité de détecter un vrai effet.

Q13. Quand utiliseriez-vous un test t vs un test de Mann-Whitney?

R13. - **t-test:** Quand les données sont approximativement normales ou n > 30 - **Mann-Whitney (non-paramétrique):** Quand la distribution n'est pas normale, données ordinales, ou petit échantillon asymétrique

---

Q14. Comment tester si le taux de défaut d'un segment (6%) diffère significativement de celui de la banque (5%)?

R14. Utiliser un **test z de proportion unilatéral**: - H_0: $p = 0.05$ - H_1: $p > 0.05$ - Calculer $z = (\hat{p} - p_0) / \sqrt{(p_0(1-p_0)/n)}$ - Comparer à la valeur critique ou obtenir la p-value

---

Q15. Qu'est-ce que le test du Chi-carré et quand l'utiliser?

R15. Le Chi-carré teste l'indépendance entre deux variables catégorielles.

H_0: Les variables sont indépendantes H_1: Les variables sont dépendantes

Exemple: Le type de produit dépend-il du segment client?

Q16. Qu'est-ce qu'ANOVA et quand l'utiliser?

R16. ANOVA (Analysis of Variance) compare les moyennes de 3+ groupes.

H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots$ (toutes les moyennes égales) H_1: Au moins une moyenne diffère

Exemple: Les soldes moyens diffèrent-ils entre les agences?

Si significatif, utiliser un test post-hoc (Tukey HSD) pour identifier les paires différentes.

Q17. Qu'est-ce qu'un intervalle de confiance à 95%?

R17. Un intervalle qui, si l'expérience était répétée de nombreuses fois, contiendrait la vraie valeur du paramètre dans 95% des cas.

IC 95% pour la moyenne: $\bar{x} \pm 1.96 \times (s/\sqrt{n})$

Attention: Ce n'est PAS "95% de chance que μ soit dans l'intervalle". μ est fixe, c'est l'intervalle qui varie.

Q18. Comment la taille d'échantillon affecte-t-elle la puissance statistique?

R18. Plus l'échantillon est grand: - Plus la puissance augmente (capacité à détecter un effet) - Plus l'erreur standard diminue - Plus les intervalles de confiance sont étroits - Plus les petits effets deviennent détectables

---

Section C: SQL (10 questions)

Q19. Quelle est la différence entre ROW_NUMBER(), RANK() et DENSE_RANK()?

R19.

```
-- Pour les valeurs: 100, 90, 90, 80
ROW_NUMBER(): 1, 2, 3, 4  -- Toujours unique
RANK():       1, 2, 2, 4  -- Saute les rangs en cas d'égalité
DENSE_RANK(): 1, 2, 2, 3  -- Ne saute pas les rangs
```

Q20. Écrivez une requête pour trouver les 3 meilleurs clients par agence en termes de solde.

R20.

```sql
WITH ranked AS (
    SELECT
        agence_id,
        client_id,
        nom,
        solde,
        ROW_NUMBER() OVER (
            PARTITION BY agence_id
            ORDER BY solde DESC
        ) as rang
    FROM clients
)
SELECT * FROM ranked WHERE rang <= 3;
```

Q21. Qu'est-ce qu'une CTE et pourquoi l'utiliser?

R21. CTE (Common Table Expression) = Requête nommée temporaire définie avec WITH.

Avantages: - Lisibilité améliorée - Réutilisation dans la même requête - Alternative aux sous-requêtes complexes - Permet la récursion

---

Q22. Expliquez le problème N+1 et comment le résoudre.

R22. Problème: 1 requête pour N éléments + N requêtes pour les détails = N+1 requêtes.

Solution: - Utiliser un JOIN - Batch loading (une requête avec IN) - Eager loading dans les ORM

Q23. Comment calculer un cumul courant en SQL?

R23.

```sql
SELECT
    date,
    montant,
    SUM(montant) OVER (ORDER BY date) as cumul
FROM transactions;
```

Q24. Quelle est la différence entre WHERE et HAVING?

R24. - **WHERE:** Filtre les lignes AVANT le GROUP BY - **HAVING:** Filtre les groupes APRÈS le GROUP BY

```sql
SELECT agence, SUM(montant) as total
FROM transactions
WHERE date >= '2024-01-01'  -- Filtre les lignes
```

```sql
GROUP BY agence
HAVING SUM(montant) > 100000;  -- Filtre les groupes
```

Q25. Comment identifier les clients inactifs depuis 90 jours?

R25.

```sql
SELECT c.*
FROM clients c
WHERE NOT EXISTS (
    SELECT 1 FROM transactions t
    WHERE t.client_id = c.client_id
    AND t.date >= CURRENT_DATE - INTERVAL '90 days'
);
```

Q26. Écrivez une requête pour calculer la variation mensuelle des ventes.

R26.

```sql
WITH monthly AS (
    SELECT
        DATE_TRUNC('month', date) as mois,
        SUM(montant) as total
    FROM ventes
    GROUP BY DATE_TRUNC('month', date)
)
SELECT
    mois,
    total,
    LAG(total) OVER (ORDER BY mois) as mois_prec,
    (total - LAG(total) OVER (ORDER BY mois)) /
        NULLIF(LAG(total) OVER (ORDER BY mois), 0) * 100 as var_pct
FROM monthly;
```

Q27. Comment optimiser une requête lente?

R27. 1. Utiliser EXPLAIN pour voir le plan d'exécution 2. Créer des index sur colonnes WHERE, JOIN, ORDER BY 3. Éviter SELECT *, sélectionner les colonnes nécessaires 4. Éviter les fonctions sur colonnes indexées dans WHERE 5. Utiliser EXISTS plutôt que IN pour les sous-requêtes 6. Limiter les résultats avec LIMIT

Q28. Quelle est la différence entre INNER JOIN et LEFT JOIN?

R28. - **INNER JOIN:** Retourne uniquement les lignes avec correspondance dans les deux tables - **LEFT JOIN:** Retourne toutes les lignes de la table gauche + correspondances de la droite (NULL si pas de correspondance)

Section D: DAX et Power BI (7 questions)

Q29. Quelle est la différence entre une colonne calculée et une mesure en DAX?

R29. | | Colonne Calculée | Mesure | |–|———————|———–| | Calcul | Au refresh | À la requête | | Stockage | Oui (mémoire) | Non | | Contexte | Ligne | Filtre | | Usage | Filtres, relations | Agrégations, KPIs |

Q30. Qu'est-ce que le contexte de filtre en DAX?

R30. Le contexte de filtre est l'ensemble des filtres actifs qui affectent un calcul: - Slicers sélectionnés - Filtres de page/rapport - Axes de visualisation - Filtres dans CALCULATE

CALCULATE modifie le contexte de filtre.

Q31. Écrivez une mesure DAX pour calculer le pourcentage du total.

R31.

```
% Total = DIVIDE(
    SUM(Ventes[Montant]),
    CALCULATE(SUM(Ventes[Montant]), ALL(Ventes))
)
```

Q32. Comment calculer un YTD (Year-to-Date) en DAX?

R32.

```
YTD Ventes = TOTALYTD(
    SUM(Ventes[Montant]),
    Calendrier[Date]
)
```

Nécessite une table de dates marquée comme calendrier.

Q33. Quelle est la différence entre ALL et ALLSELECTED?

R33. - **ALL:** Supprime TOUS les filtres de la table/colonne - **ALLSELECTED:** Supprime les filtres du visuel mais respecte les filtres du rapport (slicers externes)

Q34. Écrivez une mesure pour comparer les ventes à l'année précédente.

R34.

```
Ventes N-1 = CALCULATE(
    SUM(Ventes[Montant]),
    SAMEPERIODLASTYEAR(Calendrier[Date])
)

Var YoY =
VAR Actuel = SUM(Ventes[Montant])
VAR Precedent = [Ventes N-1]
RETURN DIVIDE(Actuel - Precedent, Precedent)
```

Q35. Pourquoi utiliser des variables (VAR) en DAX?

R35. 1. **Lisibilité:** Code plus clair et organisé 2. **Performance:** Évite de recalculer la même expression 3. **Debug:** Facilite le débogage (retourner différentes variables)

```
Mesure =
VAR Total = SUM(Ventes[Montant])
VAR Cout = SUM(Ventes[Cout])
RETURN DIVIDE(Total - Cout, Total)
```

Section E: KPIs Bancaires (5 questions)

Q36. Qu'est-ce que le NPL ratio et pourquoi est-il important?

R36. NPL Ratio = Prêts Non Performants (>90 jours) / Total des Prêts × 100

Importance: - Indicateur clé de la qualité du portefeuille de crédit - Surveillé par les régulateurs (BRH) - Impacte les provisions et la rentabilité - Benchmark: < 5%

Q37. Expliquez le CAR (Capital Adequacy Ratio).

R37. CAR = Fonds Propres Réglementaires / Actifs Pondérés par les Risques × 100

- Mesure la capacité à absorber les pertes
- Exigence BRH: ≥ 12%
- Exigence Bâle III: ≥ 8% + buffers
- Composé de Tier 1 (capital de base) et Tier 2

Q38. Quelle est la différence entre ROE et ROA?

R38. - **ROE** = Résultat Net / Capitaux Propres: Rendement pour les actionnaires - **ROA** = Résultat Net / Total Actifs: Efficacité de l'utilisation des actifs

Le ROE est généralement plus élevé car Actifs > Capitaux Propres (effet de levier). Benchmarks: ROE 12-18%, ROA 1-2%

Q39. Qu'est-ce que le ratio Loan-to-Deposit (LDR)?

R39. LDR = Total Prêts / Total Dépôts × 100

Interprétation: - < 80%: Sous-utilisation des ressources - 80-90%: Optimal - > 100%: Dépendance au financement de marché

Mesure l'équilibre entre prêts et dépôts.

Q40. Quels sont les composants du Produit Net Bancaire (PNB)?

R40.

```
PNB = Marge d'Intérêts + Commissions Nettes + Autres Revenus

Marge d'intérêts: Intérêts reçus - Intérêts payés (60-70%)
Commissions: Frais de service, cartes, transferts (20-30%)
Autres: Trading, plus-values, dividendes (5-10%)
```

Le PNB mesure la richesse créée par l'activité bancaire.

Scoring

Score	Niveau
0-15	À améliorer
16-25	Débutant
26-32	Intermédiaire
33-37	Avancé
38-40	Expert