

Fiches de Synthèse - Data Analyst UniBank Haiti

Usage: Relecture finale avant l'entretien/examen (30-45 minutes)

FICHE 1: Types de Graphiques et Usage

Graphique	Usage	Éviter si
Histogramme	Distribution continue	Peu de données
Box plot	Comparer distributions, outliers	Audience non-tech
Bar chart	Comparer catégories	Trop de catégories
Line chart	Tendances temporelles	Données non ordonnées
Scatter plot	Relations 2 variables	Variables catégorielles
Pie chart	Composition simple	> 5 catégories
Heatmap	Corrélations, patterns	Pas de pattern clair

Règle d'or: Un graphique = Un message

FICHE 2: Statistiques Descriptives

Tendance Centrale

Moyenne: $\bar{x} = \sum x / n$ → Sensible aux outliers
Médiane: Valeur centrale → Robuste aux outliers
Mode: Valeur la plus fréquente → Seul pour nominales

Dispersion

Variance: $s^2 = \sum (x - \bar{x})^2 / (n-1)$
Écart-type: $s = \sqrt{s^2}$
IQR: $Q3 - Q1$ → Pour détecter outliers
CV: $(s / \bar{x}) \times 100$ → Comparer dispersions

Forme

Skewness > 0: Queue à droite (Mode < Médiane < Moyenne)
Skewness < 0: Queue à gauche (Moyenne < Médiane < Mode)
Kurtosis > 3: Leptokurtic (queues épaisses)

FICHE 3: Tests Statistiques

Situation	Test
Moyenne vs valeur	t-test 1 échantillon
2 moyennes indépendantes	t-test 2 échantillons
2 moyennes appariées	t-test apparié

Situation	Test
3+ moyennes	ANOVA
2 proportions	z-test proportions
Indépendance catégories	Chi-carré
Non-normal, 2 groupes	Mann-Whitney
Non-normal, 3+ groupes	Kruskal-Wallis

p-value < 0.05 → Rejeter H_0 → Résultat significatif

FICHE 4: Probabilités Essentielles

Complémentaire: $P(A') = 1 - P(A)$

Union: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Intersection: $P(A \cap B) = P(A) \times P(B|A)$

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

Distributions Clés

Bernoulli: Succès/Échec, $E=p$, $Var=p(1-p)$

Binomiale: n essais, k succès, $E=np$

Poisson: Événements rares, $E=Var$

Normale: Continue, symétrique, 68-95-99.7

FICHE 5: KPIs Bancaires

Rentabilité

ROE = Résultat Net / Capitaux Propres (12-18%)

ROA = Résultat Net / Total Actifs (1-2%)

NIM = (Revenus - Charges intérêts) / Actifs productifs (3-5%)

CIR = Charges d'exploitation / PNB (45-55%)

Qualité des Actifs

NPL Ratio = Prêts > 90 jours / Total Prêts (< 5%)

Coverage = Provisions / NPL (> 100%)

Cost of Risk = Dotations provisions / Encours (1-3%)

Solvabilité & Liquidité

CAR = Fonds propres / RWA (12% BRH)

LDR = Prêts / Dépôts (80-90%)

LCR = HQLA / Sorties 30j (100%)

FICHE 6: DAX Essentiels

Agrégation

```
SUM(Table[Colonne])  
AVERAGE(Table[Colonne])  
DISTINCTCOUNT(Table[Colonne])
```

CALCULATE (Roi du DAX)

```
CALCULATE(expression, filtre1, filtre2, ...)
```

Time Intelligence

```
TOTALYTD(mesure, Calendrier[Date])  
SAMEPERIODLASTYEAR(Calendrier[Date])  
DATESINPERIOD(date, MAX(date), -12, MONTH)
```

Modificateurs

```
ALL(Table) → Supprime tous les filtres  
ALLEXCEPT(Table, Col) → Garde certains filtres  
FILTER(Table, condition) → Table filtrée
```

Pattern % du Total

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

FICHE 7: SQL Avancé

Window Functions

```
ROW_NUMBER() OVER (PARTITION BY col ORDER BY col) -- Unique  
RANK() -- Saute les rangs si égalité  
DENSE_RANK() -- Ne saute pas  
LAG(col, 1) OVER (ORDER BY date) -- Valeur précédente  
LEAD(col, 1) OVER (ORDER BY date) -- Valeur suivante
```

CTE

```
WITH cte AS (  
    SELECT ... FROM ...  
)  
SELECT * FROM cte;
```

Problème N+1

Problème: 1 requête + N requêtes supplémentaires **Solution:** JOIN ou batch loading

Optimisation

SELECT colonnes spécifiques
WHERE avec index
EXISTS plutôt que IN
SELECT *
Fonctions dans WHERE

FICHE 8: Python/Pandas Essentiels

```
# Chargement
df = pd.read_csv('file.csv')

# Exploration
df.head(), df.info(), df.describe()
df.isnull().sum()

# Filtrage
df[df['col'] > 100]
df[(cond1) & (cond2)]

# Agrégation
df.groupby('col')['val'].agg(['sum', 'mean'])

# Pivot
df.pivot_table(values='val', index='row', columns='col', aggfunc='sum')

# Valeurs manquantes
df.fillna(df['col'].median())
df.dropna(subset=['col'])
```

FICHE 9: EDA Checklist

Comprendre le contexte business
Examiner la structure (shape, dtypes, head)
Statistiques descriptives par variable
Identifier valeurs manquantes
Détecter outliers (IQR, Z-score)
Vérifier doublons
Explorer corrélations
Visualiser distributions
Documenter insights

FICHE 10: Analyse Univariée vs Multivariée

	Univariée	Bivariée	Multivariée
Variables	1	2	3+

	Univariée	Bivariée	Multivariée
Objectif	Décrire	Relation	Patterns
Outils	Stats desc, Histo	Corr, Scatter	PCA, Clustering

Corrélation (Pearson r)

$|r| < 0.3$: Faible

$0.3 \leq |r| < 0.7$: Modérée

$|r| \geq 0.7$: Forte

Corrélation \neq Causalité!

FICHE 11: Segmentation RFM

R (Recency): Jours depuis dernière activité (5=récent, 1=ancien)

F (Frequency): Nombre de transactions (5=fréquent)

M (Monetary): Montant total (5=élevé)

Segments typiques:

- Champions: R5 F5 M5
- Fidèles: R4+ F4+
- À risque: R2- F3+
- Perdus: R1 F1

FICHE 12: Intervalles de Confiance

IC 95% pour moyenne:

$$IC = \bar{x} \pm 1.96 \times (s/\sqrt{n})$$

IC 95% pour proportion:

$$IC = \hat{p} \pm 1.96 \times \sqrt{(\hat{p}(1-\hat{p}))/n}$$

Taille d'échantillon:

$$n = (z \times s / \text{marge})^2$$

FICHE 13: Indicateurs et Indices

Indicateur = Mesure simple (nb clients)

Indice = Mesure composite (indice satisfaction)

Leading indicator = Prédit l'avenir (demandes crédit)

Lagging indicator = Mesure le passé (défauts réalisés)

Stock = À un instant T

Flux = Sur une période

FICHE 14: AML Red Flags

Transactions juste sous le seuil (structuration)
Activité >> moyenne historique
Transactions avec pays à risque
Entreprises sans activité visible
Cash intensif sans justification

FICHE 15: Formules de Base

Expected Loss

$$EL = PD \times LGD \times EAD$$

Variation

$$\text{Variation \%} = (\text{Nouveau} - \text{Ancien}) / \text{Ancien} \times 100$$

CAGR (Croissance annuelle composée)

$$CAGR = (V_f/V_i)^{(1/n)} - 1$$

Checklist Jour de l'Examen

ACID: Atomicity, Consistency, Isolation, Durability
SOLID: Single, Open/Closed, Liskov, Interface, Dependency
KPIs: ROE, ROA, NPL, CAR, NIM, CIR
p-value: < 0.05 = significatif
Corrélation: -1 à +1, 0 = pas de relation linéaire
Skewness +: Queue droite, Moyenne > Médiane
NPL: > 90 jours de retard
CAR minimum BRH: 12%
CALCULATE modifie le contexte de filtre
ROW_NUMBER vs RANK vs DENSE_RANK

VOUS ÊTES PRÊT(E)! CONFIANCE!