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: x̄ = Σxᵢ / n  → Sensible aux outliers
Médiane: Valeur centrale  → Robuste aux outliers
Mode: Valeur la plus fréquente  → Seul pour nominales
```

Dispersion

```
Variance: s² = Σ(xᵢ - x̄)² / (n-1)
Écart-type: s = √(s²)
IQR: Q3 - Q1  → Pour détecter outliers
CV: (s / x̄) × 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 ∪ B) = P(A) + P(B) - P(A ∩ B)
Intersection: P(A ∩ B) = P(A) × P(B|A)
Bayes: P(A|B) = P(B|A) × 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=λ=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: Types de Variables

Classification Hiérarchique

```
QUALITATIVES (Catégorielles)
├── Nominales: pas d'ordre (type compte, région)
│   ├── Binaires: 2 catégories (oui/non)
│   └── Polytomiques: 3+ catégories
└── Ordinales: ordre naturel (rating AAA > AA > A)

QUANTITATIVES (Numériques)
├── Discrètes: entiers (nb transactions)
└── Continues: décimales (montant, taux)
```

Niveaux de Mesure

```
Nominal: = ≠ seulement
Ordinal: = ≠ < >
Intervalle: + - (zéro arbitraire)
Ratio: × ÷ (zéro absolu)
```

Statistiques par Type

```
Nominale → Mode, Chi-carré
Ordinale → Médiane, Mann-Whitney
Quantitative → Moyenne, t-test, corrélation
```

Encodage

```
Nominale → One-Hot (pd.get_dummies)
Ordinale → Label Encoding avec ordre
```

FICHE 7: SQL Avancé

Window Functions

```sql
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

```sql
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

```python
# 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 ≤ |r| < 0.7: Modérée
|r| ≥ 0.7: Forte
```

Corrélation ≠ 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 = x̄ ± 1.96 × (s/√n)

IC 95% pour proportion:
IC = p̂ ± 1.96 × √(p̂(1-p̂)/n)

Taille d'échantillon:
n = (z × s / 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 × LGD × EAD
```

Variation

```
Variation % = (Nouveau - Ancien) / Ancien × 100
```

CAGR (Croissance annuelle composée)

```
CAGR = (Vf/Vi)^(1/n) - 1
```

FICHE 16: Types de Modèles

Par Objectif

```
Descriptif: Comprendre (stats, EDA)
Prédictif: Anticiper (classification, régression)
Prescriptif: Recommander (optimisation)
```

Par Apprentissage

```
Supervisé: avec labels (défaut oui/non)
Non supervisé: sans labels (segmentation)
Semi-supervisé: mix
```

Modèles Bancaires Clés

```
Scoring crédit: Régression logistique (PD)
Fraude: Random Forest, Isolation Forest
Churn: Gradient Boosting
Segmentation: K-Means, RFM
```

Formules Risque

```
Expected Loss = PD × LGD × EAD
PD = Probability of Default
LGD = Loss Given Default
EAD = Exposure at Default
```

FICHE 17: Machine Learning Essentiels

Algorithmes Classification

```
Régression Logistique: Interprétable, scoring
Arbre de décision: Règles explicites
Random Forest: Ensemble, robuste
XGBoost: Performance maximale
```

Métriques Classification

```
Accuracy = (TP+TN) / Total
Precision = TP / (TP+FP)
Recall = TP / (TP+FN)
F1 = 2×(P×R)/(P+R)
AUC-ROC: aire sous courbe
Gini = 2×AUC - 1
```

Métriques Régression

```
MAE: erreur absolue moyenne
RMSE: racine erreur quadratique
R²: variance expliquée
```

Bonnes Pratiques

```
 Fit scaler sur TRAIN seulement
 Validation croisée
 Gérer déséquilibre classes (SMOTE)
 Data leakage
 Overfitting
```

Checklist Jour de l'Examen

```
□ Types variables: Nominale/Ordinale/Discrète/Continue
□ 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%
□ EL = PD × LGD × EAD
□ Gini = 2 × AUC - 1
□ ROW_NUMBER vs RANK vs DENSE_RANK
□ Régression logistique: odds ratio = exp(β)
□ K-Means: méthode du coude pour k
```

FICHE 18: Régression Linéaire

Modèle

$Y = \beta_0 + \beta_1 X + \varepsilon$
β_1 = Changement de Y pour 1 unité de X

Hypothèses LINE

L - Linéarité
I - Indépendance
N - Normalité résidus
E - Égalité variances (homoscédasticité)

Diagnostics

R^2: Variance expliquée (0.7 = 70%)
R^2 ajusté: Pénalise variables inutiles
VIF > 10: Multicolinéarité
Durbin-Watson ≈ 2: OK
p-value < 0.05: Coefficient significatif

FICHE 19: Séries Temporelles

Composantes TSCI

T - Tendance: Direction long terme
S - Saisonnalité: Pattern régulier
C - Cycle: Fluctuations économiques
I - Irrégulier: Bruit

Stationnarité

Test ADF: p < 0.05 → Stationnaire
Sinon: Différencier (d=1, d=2...)

Modèles

ARIMA(p,d,q): AR + Différenciation + MA
SARIMA: ARIMA + Saison
Holt-Winters: Niveau + Tendance + Saison

Métriques

MAPE < 10%: Excellent
AIC/BIC: Plus bas = Meilleur

FICHE 20: Tests Non-Paramétriques

Correspondance

```
t-test → Mann-Whitney U
t-test apparié → Wilcoxon
ANOVA → Kruskal-Wallis
Pearson → Spearman
```

Quand utiliser?

```
✓ Non normalité
✓ Petit n (< 30)
✓ Données ordinales
✓ Outliers présents
```

Corrélation Spearman

```
Basé sur les rangs
Robuste aux outliers
Relation monotone (pas linéaire)
|ρ| > 0.7 = Forte
```

FICHE 21: Cas Spéciaux Essentiels

Valeurs Manquantes

```
MCAR: Aléatoire complet → Supprimer OK
MAR: Dépend d'autres variables → Imputer par groupe
MNAR: Dépend de la valeur elle-même → Problématique

Imputation: Médiane > Moyenne (outliers)
KNN Imputer, MICE pour avancé
```

Outliers

```
IQR: Q1 - 1.5×IQR < x < Q3 + 1.5×IQR
Z-score > 3: Outlier

Traitement:
- Supprimer (avec prudence)
- Winsoriser (capper aux percentiles)
- Transformer (log)
- Flaguer (fraude!)
```

ACP

```
But: Réduire dimensions
Standardiser OBLIGATOIRE
Kaiser: Garder eigenvalues > 1
Variance cumulative > 80%
KMO > 0.6: OK pour ACP
```

ANOVA

```
3+ groupes → ANOVA (F-test)
Si significatif → Post-hoc Tukey
Levene p < 0.05 → Variances inégales → Welch ANOVA
η² > 0.14: Grand effet
```

FICHE 22: A/B Testing

Terminologie

```
Baseline: Taux actuel (contrôle)
MDE: Effet Minimal Détectable
Lift: (B - A) / A × 100%
Puissance: P(détecter vrai effet) = 80%
α: P(faux positif) = 5%
```

Étapes

```
1. HYPOTHÈSE: "B augmentera conversion de X%"
2. DESIGN: Taille échantillon, durée, métriques
3. RANDOMISER: Assignation aléatoire A/B
4. EXÉCUTER: Collecter données (≥ 7 jours)
5. ANALYSER: z-test, IC, décision
```

Pièges à éviter

```
- Peeking: Ne pas regarder avant la fin
- Multiple testing: Corriger si 3+ variantes
- Durée trop courte: Min 1 semaine
- Effet nouveauté: Peut biaiser résultats initiaux
```

Analyse

```
p < 0.05 + Lift > 0 → Déployer B
p < 0.05 + Lift < 0 → Garder A
p ≥ 0.05 → Pas de conclusion, continuer
IC ne contient pas 0 → Significatif
```

FICHE 23: Éthique et Gouvernance

Principes TERB

```
T - Transparence: Expliquer les décisions
E - Équité: Pas de discrimination
R - Responsabilité: Assumer conséquences
B - Bénéfice: Valeur pour tous
```

Biais Algorithmiques

```
Disparate Impact = Taux_minorité / Taux_majorité
DI < 0.8 (80%) → DISCRIMINATION potentielle

Variables proxy dangereuses:
- Code postal (corrélé origine)
- Prénom (corrélé genre)
```

Explicabilité (XAI)

```
SHAP: Explication locale + globale
Feature Importance: Impact de chaque variable
Droit à l'explication: Obligatoire pour refus crédit
```

Droits des Personnes (AREPO)

```
A - Accès: Voir ses données
R - Rectification: Corriger erreurs
E - Effacement: Droit à l'oubli
P - Portabilité: Récupérer ses données
O - Opposition: Refuser traitement
```

Gouvernance

```
Classification: Public < Interne < Confidentiel < Strictement confidentiel
Moindre privilège: Accès minimal nécessaire
Audit trail: Tracer tous les accès
Rétention: 10 ans transactions (légal)
```

Checklist Jour de l'Examen

```
□ Types variables: Nominale/Ordinale/Discrète/Continue
□ 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%
□ EL = PD × LGD × EAD
□ Gini = 2 × AUC - 1
□ ROW_NUMBER vs RANK vs DENSE_RANK
□ Régression logistique: odds ratio = exp(β)
□ K-Means: méthode du coude pour k
□ RÉGRESSION: LINE, VIF, DW, R²
□ SÉRIES TEMP: ARIMA, stationnarité (ADF)
□ NON-PARAM: Mann-Whitney, Kruskal-Wallis
□ MANQUANTS: MCAR/MAR/MNAR
□ OUTLIERS: IQR, winsorisation
□ A/B TEST: MDE, puissance 80%, α 5%, randomisation
□ ÉTHIQUE: Disparate Impact ≥ 0.8, SHAP explicabilité
□ DROITS: AREPO (Accès, Rectification, Effacement, Portabilité, Opposition)
```

VOUS ÊTES PRÊT(E)! CONFIANCE!