ANALYSE STATISTIQUE DES DONNÉES – EL AMAL

Nettoyage + étude statistique (dataStage.xlsx)

Oumaima Ayadi

ANNÉE UNIVERSITAIRE

2024 – 2025

PLAN DE PRÉSENTATION

1. **Introduction** : Contexte & objectifs
2. **Données** : Description & structure
3. **Nettoyage** : Préparation des données
4. **Qualité** : Vérification (NA / doublons)
5. **Statistiques** : Analyses descriptives
6. **Visualisations** : Graphiques & interprétations
7. **Conclusion** : Synthèse & recommandations

CONTEXTE & OBJECTIFS

CONTEXTE

- **Société** : EL AMAL (huilerie)
- **Source** : Données de production et transactions (Excel)
- **Période** : Données historiques de production

OBJECTIFS

- Analyser les **quantités** de production (olives, entrées)
- Étudier l'évolution des **prix** (unitaire, total)
- Identifier les **tendances** temporelles
- Produire des **visualisations claires** avec interprétation

DONNÉES UTILISÉES

SOURCE

- **Fichier** : `data/dataStage.xlsx`
- **Format** : Excel (feuille de calcul)

VARIABLES PRINCIPALES

- **Temporelles** : `date_entree` (date d'entrée)
- **Quantités** : `qte_olive`, `qte_entree` (quantités d'olives et d'entrées)
- **Prix** : `prix_total`, `prix_unitaire` (prix total et unitaire)
- **Catégorielles** : `libelle_produit`, `provenance` (produit et origine)

ÉTAPE 1 : CHARGEMENT & NETTOYAGE

OPÉRATIONS EFFECTUÉES

1. **Import** des données Excel
2. **Nettoyage** des noms de colonnes
3. **Conversion** des formats (DT, dates)
4. **Création** de variables dérivées

```
Total lignes = 32316


Lignes après filtrage = 32316
```

ÉTAPE 2 : QUALITÉ DES DONNÉES

VÉRIFICATIONS EFFECTUÉES

- **Valeurs manquantes** (NA) par variable
- **Doublons** dans le dataset
- **Cohérence** des données numériques

```
# A tibble: 1 × 2
  Variable  NA_count
  <chr>        <int>
1 cin_c_tva      246
```

```
Nombre de lignes dupliquées: 10
```

ÉTAPE 3 : STATISTIQUES DESCRIPTIVES

INDICATEURS CALCULÉS

Pour chaque variable numérique : - **Moyenne** et **médiane**
- **Écart-type**
- **Minimum** et **Maximum**

```
# A tibble: 1 × 20
  qte_olive_mean qte_olive_median qte_olive_sd qte_olive_min qte_olive_max
           <dbl>            <dbl>        <dbl>         <dbl>         <dbl>
1          1373.            1279.       12352.          15.8      1894789.
# ℹ 15 more variables: qte_entree_mean <dbl>, qte_entree_median <dbl>,
#   qte_entree_sd <dbl>, qte_entree_min <dbl>, qte_entree_max <dbl>,
#   qte_totale_mean <dbl>, qte_totale_median <dbl>, qte_totale_sd <dbl>,
#   qte_totale_min <dbl>, qte_totale_max <dbl>, prix_total_num_mean <dbl>,
#   prix_total_num_median <dbl>, prix_total_num_sd <dbl>,
#   prix_total_num_min <dbl>, prix_total_num_max <dbl>
```

ÉTAPE 4 : ANALYSE APPROFONDIE - PRIX TOTAL

STATISTIQUES DU PRIX TOTAL

Analyse détaillée de la variable `prix_total` : - Nombre d’observations
- Tendances centrales (moyenne, médiane)
- Dispersion (écart-type)
- Valeurs extrêmes

```
# A tibble: 1 × 6
  nombre moyenne mediane ecart_type minimum maximum
   <int>   <dbl>   <dbl>      <dbl>   <dbl>   <dbl>
1  32316   3230.    3000     22900.       0 3780105
```

VISUALISATION 1 : DISTRIBUTION DU PRIX TOTAL

SCATTER PLOT (NUAGE DE POINTS)

Visualisation de la **distribution** des prix totaux
Permet d’identifier les **valeurs aberrantes** et la **dispersion**

VISUALISATION 2 : RÉPARTITION DU PRIX TOTAL

HISTOGRAMME AVEC MOYENNE

Visualisation de la **distribution** des fréquences
Ligne rouge = moyenne pour référence

VISUALISATION 3 : DISTRIBUTION (ZOOM 95%)

BOX PLOT AVEC ZOOM

Visualisation de la **distribution centrale** (95% des données)
Permet d’analyser la **médiane** et les **quartiles** sans les valeurs extrêmes

---

VISUALISATION 4 : RÉPARTITION (ÉCHELLE LOGARITHMIQUE)

HISTOGRAMME EN ÉCHELLE LOG

Visualisation adaptée aux données **asymétriques**
Permet de mieux voir la distribution sur une large plage de valeurs

---

VISUALISATION 5 : ÉVOLUTION TEMPORELLE DES PRIX

PRIX MOYEN MENSUEL

Analyse de l'**évolution** du prix moyen par mois
Identification des **tendances** et **saisonnalités**

VISUALISATION 6 : ANALYSE PAR PRODUIT

DISTRIBUTION DES PRIX (TOP 5 PRODUITS)

Comparaison des **prix** entre les **5 produits les plus fréquents**
Identification des **différences** de prix entre produits

VISUALISATION 7 : RELATION QUANTITÉ/PRIX

SCATTER PLOT AVEC RÉGRESSION LINÉAIRE

Analyse de la **corrélation** entre quantité et prix
Ligne noire = tendance linéaire (régression)

VISUALISATION 8 : ÉVOLUTION DES QUANTITÉS

SÉRIE TEMPORELLE DES QUANTITÉS

Analyse de l'**évolution** des quantités totales dans le temps
Identification des **périodes** de forte/faible production

VISUALISATION 9 : ÉVOLUTION DES PRIX TOTAUX

SÉRIE TEMPORELLE DES PRIX

Analyse de l'**évolution** des prix totaux dans le temps
Identification des **tendances** de prix

VISUALISATION 10 : DISTRIBUTION GÉNÉRALE DES PRIX

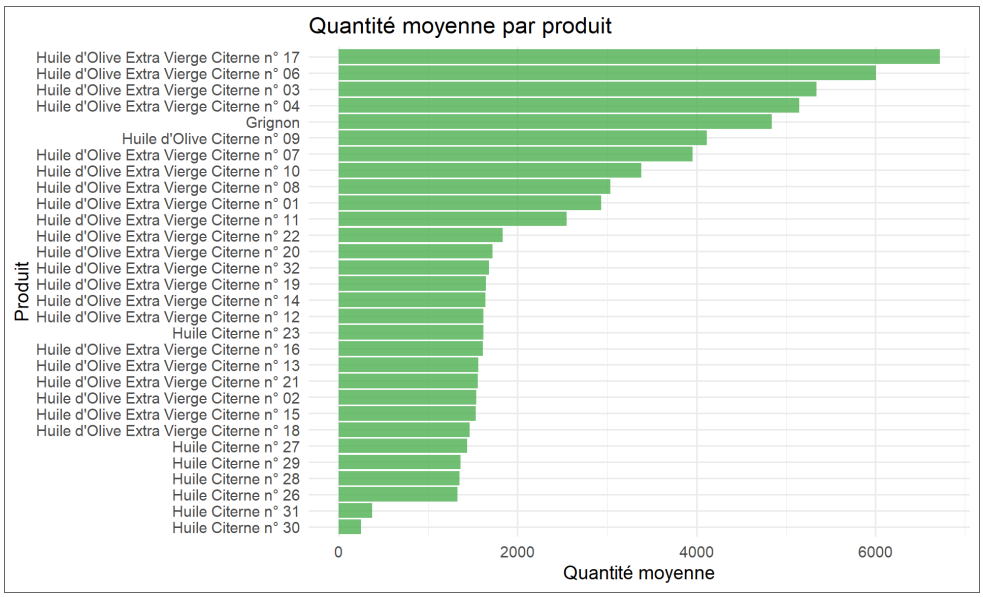
HISTOGRAMME DES PRIX

Vue d’ensemble de la **distribution** des prix totaux
Identification de la **forme** de la distribution

VISUALISATION 11 : QUANTITÉS PAR PRODUIT

BARRES HORIZONTALES

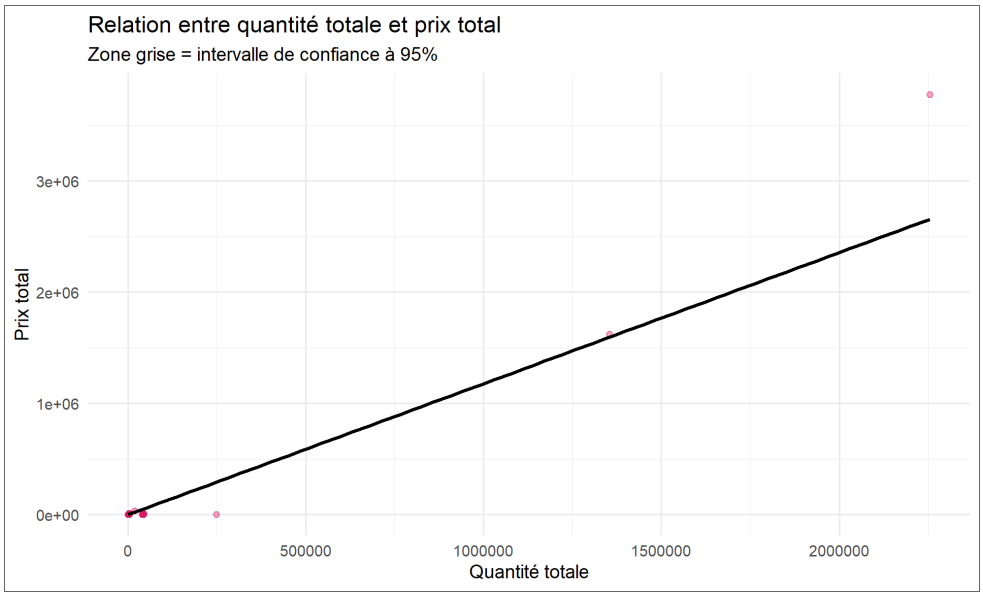
Comparaison des **quantités moyennes** par produit
Identification des produits les plus **produits**



VISUALISATION 12 : RELATION PRIX/ QUANTITÉ (DÉTAILLÉE)

SCATTER PLOT AVEC INTERVALLE DE CONFIANCE

Analyse approfondie de la **relation** quantité/prix
Zone grise = intervalle de confiance de la régression



SYNTHÈSE DES RÉSULTATS

POINTS CLÉS IDENTIFIÉS

- **Nettoyage** : Conversion réussie des formats (DT, dates)
- **Qualité** : Données vérifiées (NA, doublons)
- **Statistiques** : Indicateurs descriptifs calculés
- **Visualisations** : 12 graphiques produits pour l’analyse

INTERPRÉTATIONS PRINCIPALES

DISTRIBUTION DES PRIX

- Distribution **asymétrique** (beaucoup de petits prix, quelques très grands)
- Présence de **valeurs aberrantes** à traiter
- Prix moyen vs médiane : indicateur d'**asymétrie**

ÉVOLUTION TEMPORELLE

- **Tendances** identifiables dans les séries temporelles
- **Saisonnalité** possible à analyser plus en détail
- **Corrélation** entre quantité et prix observée

CONCLUSION & RECOMMANDATIONS

RÉALISATIONS

Nettoyage complet des données effectué
Statistiques descriptives calculées
Visualisations claires et interprétables produites
Relations entre variables identifiées

RECOMMANDATIONS

1. **Traitement des valeurs aberrantes** : Examiner les prix très élevés
2. **Analyse de saisonnalité** : Étudier les variations mensuelles/annuelles
3. **Segmentation par produit** : Approfondir l'analyse par type de produit
4. **Modélisation** : Développer des modèles prédictifs si nécessaire

PERSPECTIVES

- Analyse de **régression** plus poussée
- **Tests statistiques** (normalité, corrélations)
- **Clustering** des produits par prix/quantité

Speaker notes