2025 - 2026

Loan THOMY
Nexa Digital School

Étude de cas - Bloc 2

1. Reformulation du besoin

SmartMarket est un e-commerçant spécialisé dans les accessoires technologiques et a mené simultanément plusieurs campagnes marketing (Emailing, Facebook / Instagram Ads, LinkedIn Ads, actions CRM (Customer Relationship Management).
La direction marketing souhaite identifier les canaux les plus performants, les profils et segments les plus réactifs et des pistes d'optimisation des investissements.
Le périmètre imposé couvre uniquement les leads de septembre 2025 et les KPI (Key Performance Indicator) prioritaires attendus sont : le CTR (Click-Through Rate), le taux de conversion et le coût par lead, ainsi qu'une lecture par segments en fonction de la région, des profils CRM et de l'appareil.

2. Données disponibles et périmètre d'analyse

2.1 Sources de données

Nous avons 3 sources de données fournies, un CSV de leads (lead_id, date, channel et device), un JSON de performance campagne (channel, cost, impressions, clicks et conversions) et un fichier CRM (lead_id, company_size, sector, region, status).

Ces sources couvrent deux granularités, le lead-level (leads + CRM) et le channel-level qui correspond aux campagnes. Ce qui impose une vigilance sur les jointures et l'interprétation.

2.2 Filtrage du périmètre (septembre 2025)

Le périmètre temporel retenu est du 01-09-2025 au 30-09-2025, conformément à la note explicative et toutes les observations en dehors de cette période sont donc exclues.
Les canaux retenus sont ceux présents dans les datasets fournis (Emailing, Facebook Ads, Instagram Ads, LinkedIn) car l'objectif porte sur la comparaison inter-canaux et l'optimisation des investissements par canal.

3. Sélection du périmètre et des variables

1.1 - Choix des variables gardées

Pour mesurer la **performance marketing** (KPI), les variables indispensables sont cost, impressions, clicks, conversions, channel, car elles permettent de calculer le CTR, le taux de conversion, le CPC (coût par clic), le CPL (coût par lead) et de comparer les canaux.

Pour analyser les **profils** et la **réactivité,** les variables nécessaires sont : company_size, sector, region, status, device et channel au niveau lead afin de relier segments et les comportements (distribution …)
Pour le **périmètre** et le **contrôle qualité**, nous avons besoin de conserver lead_id et date car le lead_id sert de clé de jointure et de déduplication, date sert au filtrage temporel

1.2 - Choix des variables non-gardées

Les variables non liées à la performance, aux segments ou au périmètre seraient exclues afin de respecter la minimisation, ce qui est également utile pour le RGPD et afin d'éviter le bruit analytique.
Les doublons éventuels sur lead_id sont supprimés en conservant l'observation la plus récente car un lead doit être unique dans une analyse de funnel et de segmentation.

4. Choix du SGBD

4.1 - SGBD retenu

Le SGBD retenu est SQLite car l'analyse local (POC) et SQLite offre un stockage relationnel simple sans serveur et adapté aux requêtes de synthèses (groupby, agrégations).
SQLite permet également de documenter un schéma, des index et une comparaison de temps d'exécution sur des requêtes de segmentation.

4.2 - Alternatives

PostgreSQL aurait été recommandé dans un contexte de production, mais sur un exercice en local avec peu de données, SQLite est plus adapté et moins complexe à mettre en place.

5. Intégration de données (ETL)

5.1 - Etapes réalisées

1. Ingestion : lecture du CSV leads, JSON campagnes et du XLSX CRM
2. Nettoyage : typage des dates, contrôle des numériques, suppression des doublons, normalisation de libellés
3. Filtrage périmètre : conservation des leads de septembre 2025
4. Jointures : leads et CRM via lead_id (one to one) et enrichissement par KPI campagne via channel (many to one)
5. Production : export d'une table analytique + export de tables KPI + export des visualisations créées

5.2 - Points de vigilance entre granularité canal et lead

Le dataset ne fournit pas de lien direct entre un lead et une campagne précise donc l'enrichissement KPI se fait au niveau canal, cela doit être documenté sinon il y a un risque de sur-interprétation.

6. Gestion des valeurs manquantes et contrôle qualité

6.1 Règles appliquées

Les champs structurants (date, lead_id, channel) sont considérés obligatoires pour les leads, donc les lignes où ils manquent sont exclues car elles ne sont ni filtrables ni exploitables analytiquement.

Pour le CRM, la jointure est faite en left join afin de ne pas exclure un lead qui n'aurait pas d'attribut CRM car on préfère garder le périmètre et rendre le manque explicite, puis traiter / compléter si besoin.

Pour les campagnes, les lignes incohérentes (impressions ≤ 0, valeurs numériques non typables) sont supprimées car elles empêchent le calcul fiable de CTR et du taux de conversion.

6.2 Contrôles et traçabilité

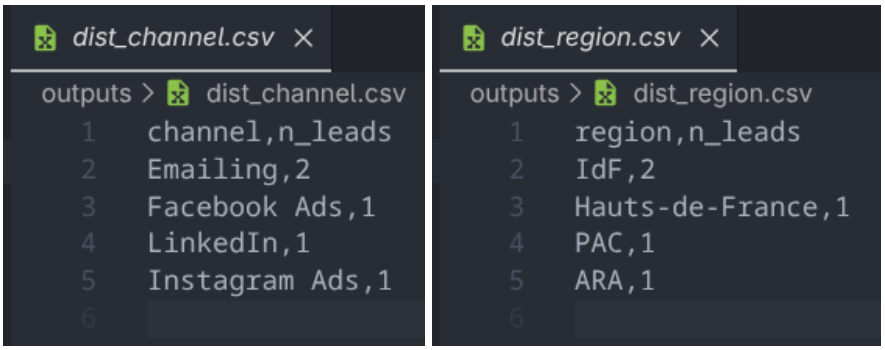
Les tables intermédiaires et finales sont exportées en CSV (table lead enrichie, KPI campagnes, distributions et crosstabs), ce qui garantit la traçabilité et la reproductibilité de l'analyse (audit possible).

7. Analyse statistique univariée et bivariée

7.1 Analyse univariée

Les distributions qualitatives sont calculées sur : canal, device, région, statut, secteur afin d'obtenir fréquences, proportions et dominantes de population.

Les statistiques quantitatives descriptives sont produites sur les métriques campagne et les KPI dérivés (min, max, moyenne, dispersion via describe).
(voir ci-dessous : dist_channel.csv, dist_region.csv, campaign_quant_summary.csv)



campaign_quant_summary.csv

outputs > campaign_quant_summary.csv

```
,cost,impressions,clicks,conversions,ctr,conv_rate_click,cpc,cpl
count,4.0,4.0,4.0,4.0,4.0,4.0,4.0,4.0
mean,2125.0,62500.0,1800.0,146.25,0.0285218253968254,0.08148018648018647,1.4084304584304583,17.142857142857142
std,788.986691902975,22173.55782608345,752.772652709081,60.466933112239126,0.006567460317460317,0.0014427755827919578,0.9591
min,1200.0,40000.0,900.0,75.0,0.0225,0.08,0.6923076923076923,8.571428571428571
25%,1650.0,47500.0,1350.0,108.75,0.023958333333333335,0.08057692307692307,0.7730769230769231,9.642857142857142
50%,2150.0,60000.0,1850.0,150.0,0.027222222222222224,0.0812937062937063,1.0818181818181818,13.333333333333334
75%,2625.0,75000.0,2300.0,187.5,0.031785714285714285,0.0821969696969697,1.7171717171717171,20.833333333333336
max,3000.0,90000.0,2600.0,210.0,0.037142857142857144,0.08333333333333333,2.7777777777777777,33.333333333333336
```

7.2 Analyse bivariée

Des croisements “métier” sont produits : canal x statut (proxy qualité), région x statut (priorisation commerciale/CRM), et une table de comparaison KPI campagne (CTR, CPL, taux conv) pour arbitrage budgétaire.

(Voir ci-dessous : ctab_status_by_channel.csv, ctab_status_by_region.csv, campaign_perf_table.csv)

ctab_status_by_channel.csv

outputs > ctab_status_by_channel.csv

```
channel,CLient,Client,MQL,SQL,All
Emailing,0,0,2,0,2
Facebook Ads,0,0,0,1,1
Instagram Ads,0,1,0,0,1
LinkedIn,1,0,0,0,1
All,1,1,2,1,5
```

ctab_status_by_regions.csv

outputs > ctab_status_by_regions.csv

```
channel,CLient,Client,MQL,SQL,All
Emailing,0,0,2,0,2
Facebook Ads,0,0,0,1,1
Instagram Ads,0,1,0,0,1
LinkedIn,1,0,0,0,1
All,1,1,2,1,5
```

campaign_perf_table.csv

outputs > campaign_perf_table.csv

```
channel,ctr,cpl,conv_rate_click
Instagram Ads,0.037142857142857144,8.571428571428571,0.08076923076923077
Emailing,0.03,10.0,0.08
Facebook Ads,0.024444444444444446,16.666666666666668,0.08181818181818182
LinkedIn,0.0225,33.333333333333336,0.08333333333333333
```

7.3 Interprétations attendues

Le canal avec le CTR le plus élevé (voir campaigns_kpi.csv) indique une meilleure capacité à générer des clics à partir des impressions, ce qui est utile pour juger l’attractivité créative / ciblage.

Le canal avec le CPL le plus faible indique une meilleure efficience budgétaire pour obtenir une conversion (lead), ce qui est un critère clé d’optimisation d’investissement.

Le croisement canal x statut permet d’éviter un biais “volume only” et d’orienter vers des canaux potentiellement plus qualifiants.

Résultats (voir ci-dessous campaigns_kpi.csv) :

campaigns_kpi.csv

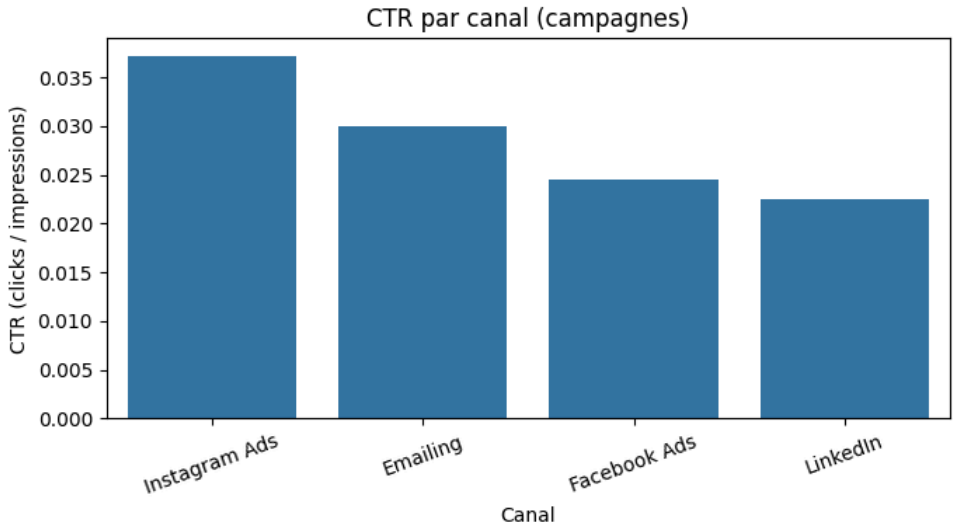
outputs > campaigns_kpi.csv

```
campaign_id,channel,cost,impressions,clicks,conversions,ctr,conv_rate_click,conv_rate_impr,cpc,cpl
CAMP01,Emailing,1200,50000,1500,120,0.03,0.08,0.0024,0.8,10.0
CAMP02,Facebook Ads,3000,90000,2200,180,0.024444444444444446,0.08181818181818182,0.002,1.3636363636363635,16.666666666666668
CAMP03,LinkedIn,2500,40000,900,75,0.0225,0.08333333333333333,0.001875,2.7777777777777777,33.333333333333336
CAMP04,Instagram Ads,1800,70000,2600,210,0.037142857142857144,0.08076923076923077,0.003,0.6923076923076923,8.571428571428571
```

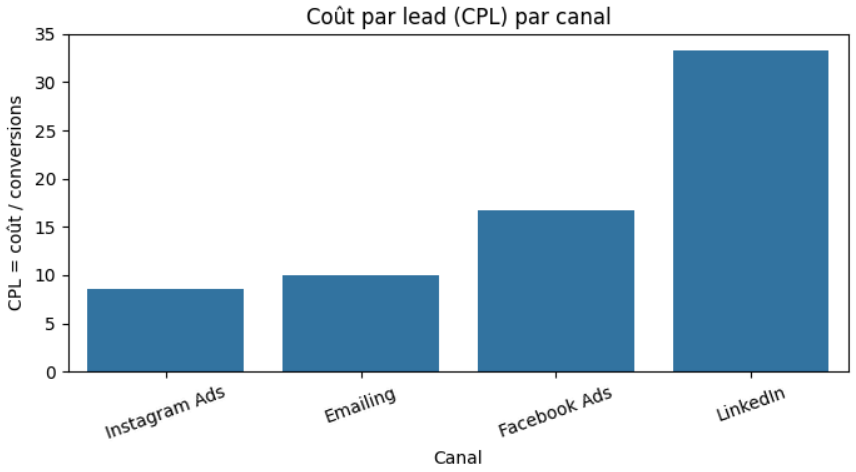
8. Visualisations

Nous allons maintenant voir les visualisations produites qui répondent chacune à une question métier explicite (performance canal, coût, trade-off, segmentation, qualité), avec titres, axes lisibles et légendes.

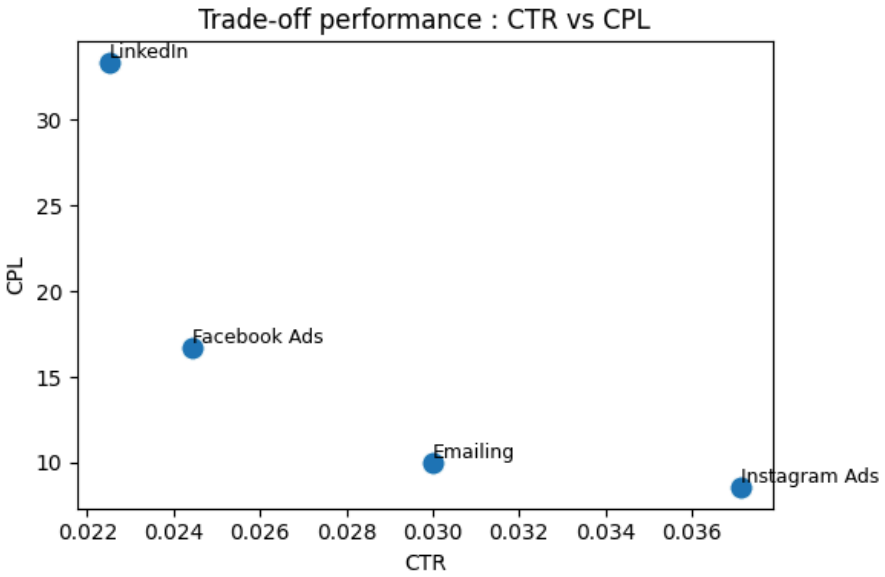
8.1 - Graphique 1 : CTR par canal



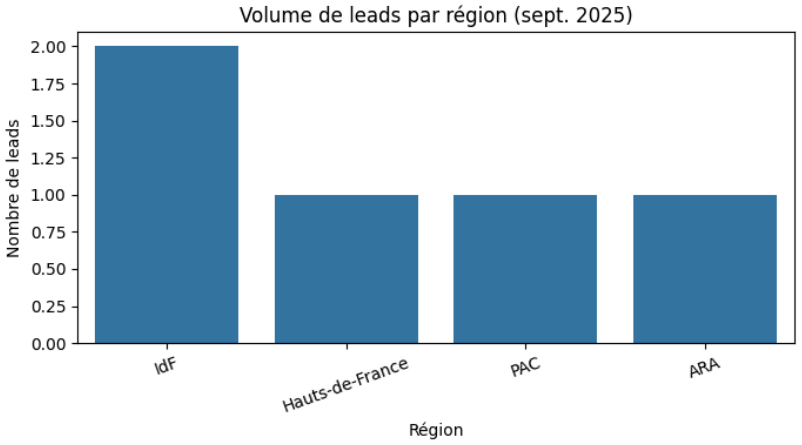
8.2 - Graphique 2 : CTL par canal



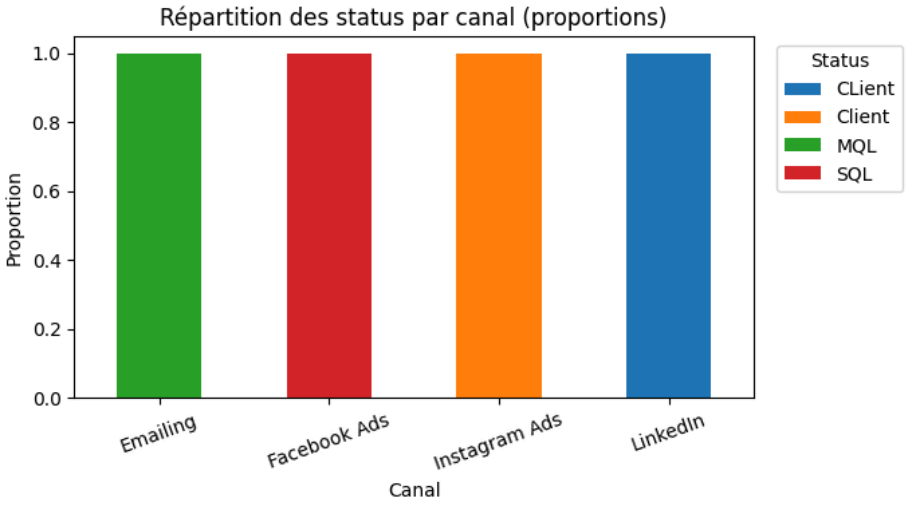
8.3 - Graphique 3 : Trade-off CTR vs CPL



8.4 - Graphique 4 : Volume de leads par région



8.4 - Graphique 5 : Répartition des statuts par canal



9. Tableau de bord synthétique

9.1 Objectif du dashboard

Le dashboard vise une lecture “direction marketing” : 3 à 6 KPI maximum + quelques graphiques clés pour décider rapidement sur l’allocation budgétaire et les priorités de ciblage.

9.2 KPI retenus

- Nombre de leads (périmètre septembre 2025)
- Coût total campagnes
- CTR global
- Taux de conversion global (sur clic)
- CPL global

Valeurs exactes (Voir ci-dessous global_kpis.csv) :

global_kpis.csv

outputs > global_kpis.csv

```
1  n_leads_sept_2025,total_cost,global_ctr,global_conv_rate_click,global_cpl
2  5,8500.0,0.0288,0.08125,14.52991452991453
```

10. Schéma de base de données

Voici le schéma de la base donnée :

- Table leads_enriched : lead_id (PK), date, channel, device, company_size, sector, region, status, + colonnes KPI canal (cost, impressions, clicks, conversions).

- Table campaigns_kpi : channel (PK), campaign_id, métriques brutes + KPI dérivés (ctr, cpl, cpc, etc.).

Ce schéma permet de requêter la segmentation sur leads_enriched et de comparer la performance canaux sur campaigns_kpi.

En réalité, je n'ai pas eu le temps de développer la base de données avec SQLite dans le code fourni, j'ai seulement effectué les comparaisons de temps d'exécution sans les sauvegarder en base de données (voir le prochain chapitre).

11. Comparaison des temps d'exécution

Comme expliqué précédemment, je n'ai pas eu le temps de développer la base de données avec SQLite.

Résultats (voir perf_comparaison.csv et le code associé ci-dessous) :

```python
# Timing comparisons

def build_perf_comparaison(lead_tbl, out_path, scale_rows=200_000, n_runs=20):
    base = lead_tbl[["channel", "status"]].copy()

    reps = max(1, scale_rows // max(1, len(base)))
    big = pd.concat([base] * reps, ignore_index=True)

    def time_groupby(df, label):
        t0 = time.perf_counter()
        for _ in range(n_runs):
            _ = df.groupby(["channel", "status"], sort=False, observed=True).size()
        t1 = time.perf_counter()
        return (t1 - t0) / n_runs

    big_obj = big.copy()
    t_obj = time_groupby(big_obj, "baseline_object")

    big_cat = big.copy()
    big_cat["channel"] = big_cat["channel"].astype("category")
    big_cat["status"] = big_cat["status"].astype("category")
    t_cap = time_groupby(big_cat, "optimized_category")

    perf = pd.DataFrame(
        [
            {"method": "baseline_object", "rows": len(big_obj), "avg_time_s": t_obj},
            {"method": "optimized_category", "rows": len(big_cat), "avg_time_s": t_cap},
        ]
    )
    perf["speedup"] = perf["avg_time_s"].iloc[0] / perf["avg_time_s"]
    perf.to_csv(out_path, index=False)
    return perf
```

perf_comparaison.csv

outputs > perf_comparaison.csv

```
method,rows,avg_time_s,speedup
baseline_object,200000,0.015876265649967535,1.0
optimized_category,200000,0.008565660050044244,1.8534783726194608
```

12. Points RGPD et éthique

12.1 - Application du RGPD

Voici les points à appliquer pour se conformer au RGPD cette étude de cas :

- La minimisation des données doit être appliquée en ne conservant que les variables nécessaires aux KPI et à la segmentation attendue, ce qui réduit l'exposition RGPD et évite de collecter des attributs non utiles.

- Les identifiants doivent être pseudonymisés et la donnée personnelle (dans le cas où elle existe : email, nom, téléphone) ne devrait pas être intégrée à ce stade si elle n'est pas strictement nécessaire aux analyses demandées.

12.2 - Intégration de l'éthique

Concernant l'éthique, la segmentation par région/secteur/taille doit être utilisée à des fins d'optimisation marketing sans générer d'exclusion illégitime ou de discrimination et la transparence sur la finalité (mesure performance, personnalisation) doit être obligatoirement assurée.

13. Conclusion

Pour conclure, la comparaison inter-canaux via CTR, taux de conversion et CPL nous permet d'identifier les canaux à privilégier budgétairement (efficience) et ceux à optimiser (trade-off CTR élevé mais CPL trop haut, ou inversement).

Enfin, Les crosstabs canal x statut et région x statut nous permettent de passer d'une lecture marketing à une lecture qualité commerciale et de proposer des actions concrètes : ciblage, A/B tests et de prioriser les CRM / commerciale sur des segments plus porteurs.