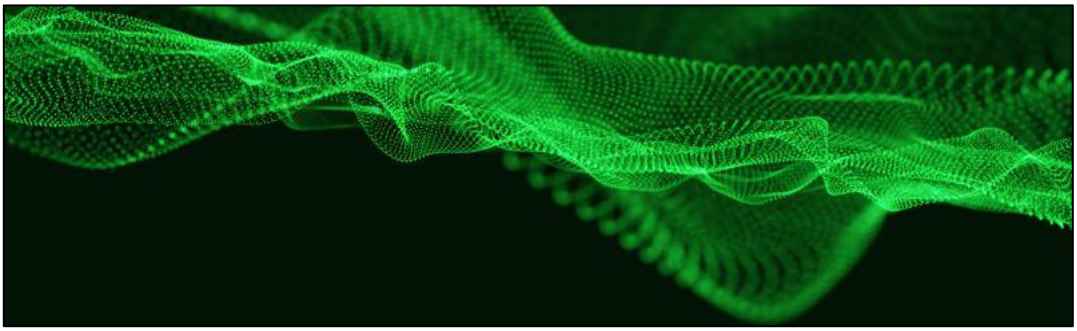
Data Cat Ink Company



Contenu

I. Présentation de l'entreprise

À propos de Data Cat Inc :

Data Cat Inc est une entreprise spécialisée dans l'analyse et la gestion des données. Fournissant des services variés allant de la collecte de données à l'analyse avancée, elle aide ses clients à transformer leurs données en outils stratégiques permettant de maximiser la performance et l'efficacité opérationnelle.

Client : **Green Bikes Tours**

Green Bikes Tours est une compagnie de tours guidés qui opère principalement dans la ville de Paris. Green Bikes Tours propose quatre types de circuits quotidiens, dont le plus populaire est le Night Bike Tour.

II. Cadre et contexte de l'analyse

Période : 28 jours, du 23-05-2022 au 19-06-2022

Données disponibles : 9 fichiers de type plat, chacun contenant des informations spécifiques (e.g., point de contrôle, enregistrements de caisse, réservations, informations clients, météo).

Moyens techniques : Python (data-wrangling), Power BI pour visualisation, et Jupyter Notebook pour l'analyse et le développement des scripts.

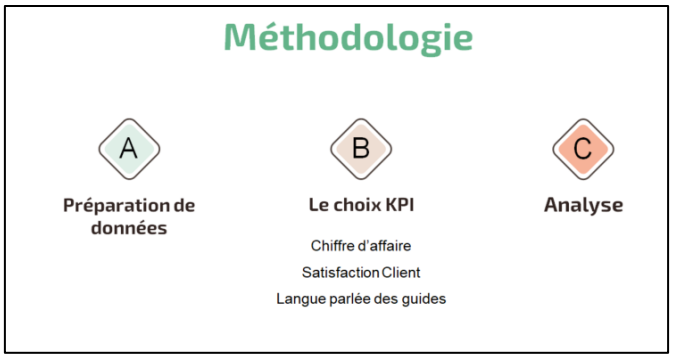
Temps alloué : 5 jours

Contexte de l'analyse : Pour maximiser leur performance, Green Bikes Tours a fait appel à Data Cat Inc afin d'optimiser le Night Bike Tour. L'objectif est de mieux comprendre les dynamiques qui influencent le succès de cette activité phare et d'identifier des axes d'amélioration. En donnant accès à un extrait de données de 28 jours (du 23-05-2022 au 19-06-2022), Green Bikes Tours souhaite évaluer les capacités analytiques de Data Cat Inc pour potentiellement renforcer cette collaboration

Objectifs

- Amélioration de l'activité : Identifier les axes d'optimisation du Night Bike Tour afin d'accroître le chiffre d'affaires et l'efficacité opérationnelle.
- Influence de la météo : Analyser l'impact des conditions météorologiques sur le chiffre d'affaires et la participation au Night Bike Tour.
- Satisfaction client : Évaluer les niveaux de satisfaction client et proposer des actions d'amélioration.
- Compétence linguistique des guides : Vérifier si les guides maîtrisent l'anglais pour répondre aux attentes des clients internationaux.

III. Méthodologie



Étapes de l'analyse *: Préparation des données, choix des KPI, analyse*

Préparation des données :

Consolidation et nettoyage des fichiers plats, identification des éventuelles incohérences, et transformation des données pour une analyse plus fluide.

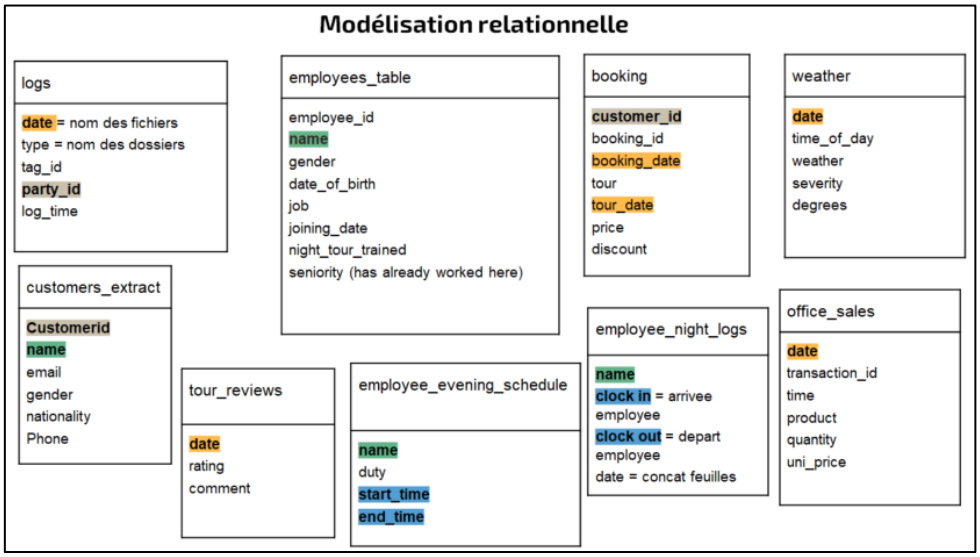
Choix des KPI :

Chiffre d'affaires : Suivi des revenus générés pour chaque jour et par influence de la météo.

Satisfaction client : Analyse des retours clients et des avis pour évaluer l'expérience.

Langue parlée par les guides : Mesurer le niveau de compétence des guides en anglais pour aligner les offres sur les besoins des clients internationaux.

Modélisation :



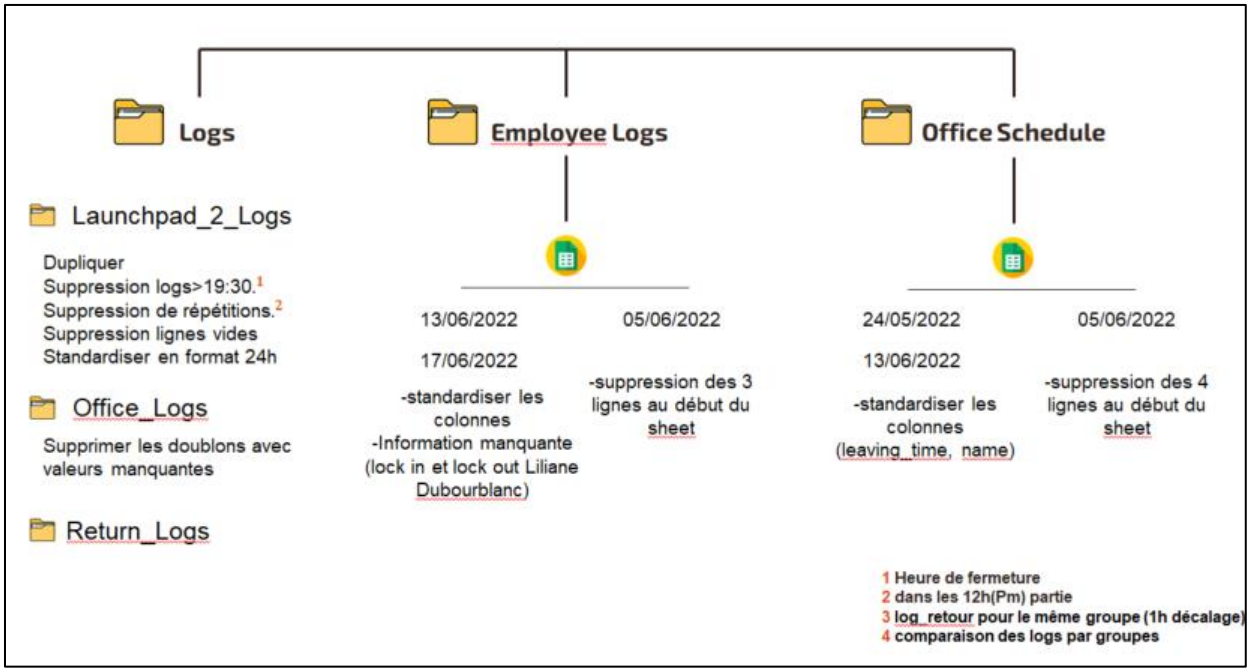
Le modèle de données *: modélisation, cardinalités*

Traitement de qualité de données :

Qualité des données :

Qualité de données		
Dossier	**Données manquantes**	**Doublons**
Logs Launchpad 2 Office Logs Return Logs	 0,4 % 0,3 % 0,1 %	 3,51 % 4,47 % 8,14 %
Employee Logs	Liliane Dubourblanc	10,87 %
Tour Manifest	Dossier non fourni	
Standardisation des formats	Date : datetime log_time : datetime No.Poeple : duty Name : name	

Nettoyage des données



→ Colonne calculée : chiffre
→ Standardisation des dates dans date et log_time =>

- Standardisation de format Date : en AAAA-MM-JJ
- Standardisation de format Heure : en h:m:s (format 24H)

→ Remplacer Name par name
→ Suppression des doublons
→ Nettoyage :

Logs :

- launchpad2 :

- suppression des lignes log_time > 19h30 (fermeture du bureau)
- en cas de doublon de tag_id, nous gardons le plus petit des 2 car c'est celui le plus proche du log_time de count_and_send

- return_logs :

- suppression des log_time > 23h30 (fin du tour des guides : 23h, fermeture du bureau : 23h30)
- en cas de doublon du tag_id, nous gardons le log-time le plus grand car les groupes rentrent ensemble

employee_logs :

- suppression de « Lilianne Dubourblanc » car donnée manquante

Office schedule :

- 24-05-2022 -> suppression de la colonne « leaving_time »

Analyse des KPIs :

Calcule du chiffre d'affaires :

- CA = Prix de réservation + total des ventes
  - Prix par réservation = Nombre de réservation pas personne x prix unitaire (50)
  - Total des ventes = Prix de vente par produit x quantité

IV. Explication Technique et Outils Utilisés

Jupyter notebook (python) : Pandas,

Fonctions : pd.read_csv(), df.head(), df.info(), pd.to_datetime(), df.to_csv(), os.listdir(), pd.concat(), reset_index(), drop(), rename(), to_excel(), isnull().sum(), duplicated().sum(), sort_values(), pd.read_excel().

Power BI : visualisation

Description Script python :

- ***Employee_clean*** :

Objectif : charger et nettoyer les fichiers journaliers de logs d'activités, puis de les concaténer en un seul DataFrame nettoyé. Le script gère les doublons et les valeurs manquantes pour préparer les données pour l'analyse.

Chargement des Données : Charge les données des employés à partir d'un fichier CSV (employees_table.csv) en utilisant un point-virgule comme séparateur.

Affichage Initial des Données : Affiche les cinq premières lignes du DataFrame pour un aperçu rapide des données.

Informations Générales : Fournit des informations structurelles sur le DataFrame, y compris les types de données et le nombre de valeurs non nulles.

Vérification des Doublons : Vérifie la présence de lignes en double dans le DataFrame et spécifiquement dans la colonne employee_id.

Conversion de Dates : Convertit les colonnes date_of_birth et joining_date au format date.

Création de Noms : Crée une nouvelle colonne name en combinant employee_first_name et employee_last_name.

Suppression de Colonnes : Supprime les colonnes jugées inutiles du DataFrame.

Exportation des Données : Enregistre le DataFrame nettoyé dans un nouveau fichier CSV nommé employee_clean.csv, sans inclure l'index des lignes.

- ***Logs_lean :***

Objectifs : fusionner les horaires de travail des employés provenant de plusieurs feuilles d'un fichier Excel en un seul jeu de données structuré. Le script ajoute des colonnes de dates, nettoie les données et vérifie la qualité avant de les exporter pour une analyse complète du planning des employés

Chargement des Fichiers : Liste les fichiers dans le dossier Logs/Count_Send_Logs et charge chaque fichier CSV dans un dictionnaire, où la clé est une partie du nom du fichier.

Concaténation des Données : Combine tous les DataFrames stockés dans le dictionnaire en un seul DataFrame (concat_df) et réinitialise l'index.

Suppression de Colonnes : Supprime des colonnes non pertinentes (index, level_1) et renomme une colonne (level_0 en date), tout en convertissant les dates au format datetime.

Exportation vers Excel : Enregistre le DataFrame concaténé dans un fichier Excel, nommé selon le dossier d'entrée.

Vérification des Données : Pour plusieurs fichiers Excel chargés, vérifie les valeurs nulles et les doublons, et analyse la validité des données (par exemple, en se basant sur un maximum d'enregistrements par Launchpad).

Traitement des Doublons : Localise et gère les doublons par rapport à tag_id, en conservant les enregistrements avec le log_time le plus proche.

Filtrage des Données : Filtre les enregistrements en fonction de critères spécifiques, notamment des horaires pour log_time

- ***Office_schedule :***

Objectif : Consolider les données de planning des employés provenant de plusieurs feuilles et fichiers, de les nettoyer, puis de les enregistrer en un seul jeu de données pour une analyse ultérieure.

Chargement et Combinaison des Données : Charger toutes les feuilles d'un fichier Excel principal, ajouter une colonne de date à chaque DataFrame, puis les concaténer en un seul.

Transformation des Données : Charger des feuilles spécifiques individuellement, ajuster les noms de colonnes et supprimer les colonnes inutiles.

Agrégation Finale des Données : Fusionner ces fichiers individuels et nettoyer les colonnes supplémentaires sans nom.

Exportation et Vérification de Qualité : Enregistrer les données nettoyées finales sous forme de fichier CSV, puis vérifier les valeurs manquantes et les doublons pour assurer la qualité des données.

V. Visualisation des Résultats

Après avoir réalisé les scripts Python et préparé les données, je les ai téléchargées dans Power BI pour illustrer les KPI clés. Dans ce tableau de bord, j'ai visualisé plusieurs indicateurs liés au chiffre d'affaires.

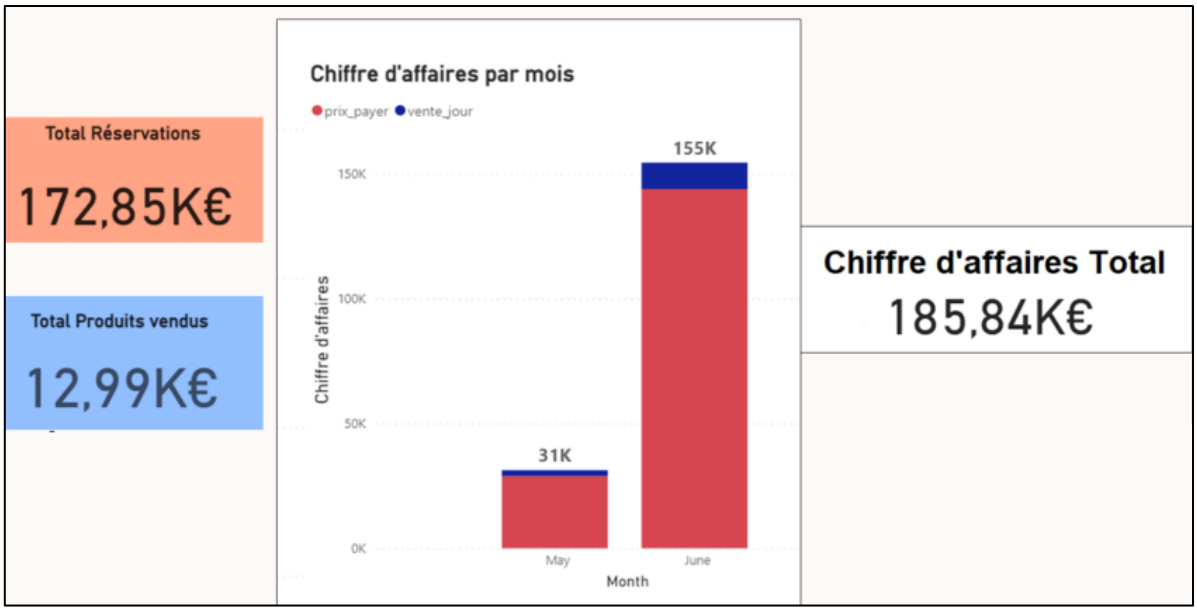
KPI-Chiffre d'Affaires

→ Le **Chiffre d'Affaires :** Le chiffre d'affaires est principalement basé sur le prix_payer, comme le montre le graphique en barres, avec une contribution supplémentaire de la vente_jour (en bleu). Ce graphique présente le chiffre d'affaires par mois (mai et juin) et met en évidence une nette augmentation de mai (31K€) à juin (155K€).

Sur la gauche, les indicateurs de performance détaillent :

- Total Réservations : 172,85K€, montrant le revenu généré par toutes les réservations.
- Total Produits vendus : 12,99K€, représentant le revenu lié à la vente directe de produits.

Enfin, le chiffre d'affaires total est de 185,84K€, indiquant le montant total consolidé pour la période.



kpi : l'Évolution du Chiffre d'Affaires Mensuel

KPI-Best seller

→ On observe que le T-shirt est de loin le produit le plus vendu, générant plus de 6K€. Il est suivi par d'autres articles, comme les petites bouteilles d'eau, les grandes bouteilles d'eau, les chips et les porte-clés, mais leur contribution est bien moindre

→ Le chiffre d'affaires des réservations par jour de la semaine. Les revenus augmentent progressivement au fil de la semaine, avec un pic le dimanche, où le chiffre d'affaires atteint environ 35K€, indiquant que les réservations sont particulièrement élevées pendant le week-end

Comparaison : *les bests-ellers de l'entreprise*

KPI-Influence météorologique

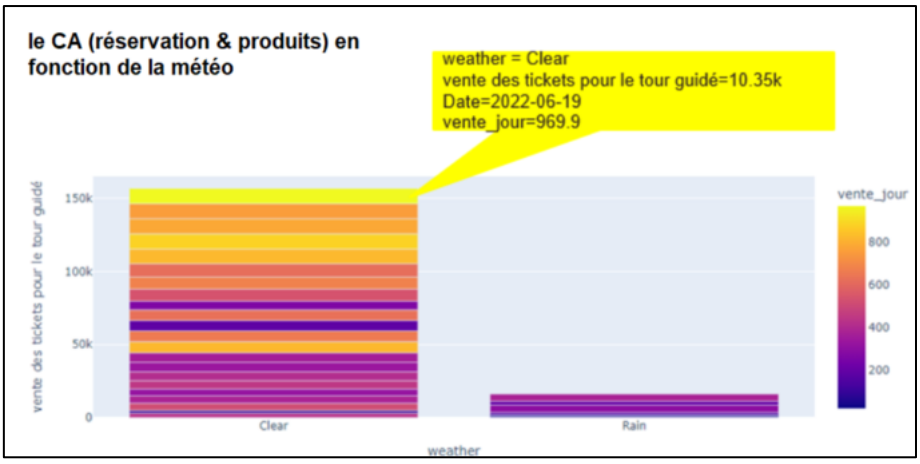
L'impact significatif des conditions météorologiques sur le chiffre d'affaires généré par la vente de billets pour les visites guidées.

Observations clés :

→ ***Corrélation positive forte***: Il existe une corrélation positive très marquée entre le nombre de jours ensoleillés et le volume des ventes. Les journées ensoleillées sont associées à une augmentation significative du chiffre d'affaires.

→ ***Effet saisonnier potentiel***: Bien que le graphique ne présente pas de données sur une longue période, il est possible qu'un effet saisonnier vienne s'ajouter à l'impact de la météo. Par exemple, les mois d'été, généralement plus ensoleillés, pourraient générer des ventes plus élevées en raison des vacances scolaires.

→ ***Segmentation des clients***: Il serait intéressant d'analyser si les différents profils de clients (familles, couples, groupes d'amis) réagissent de manière différente aux conditions météorologiques.

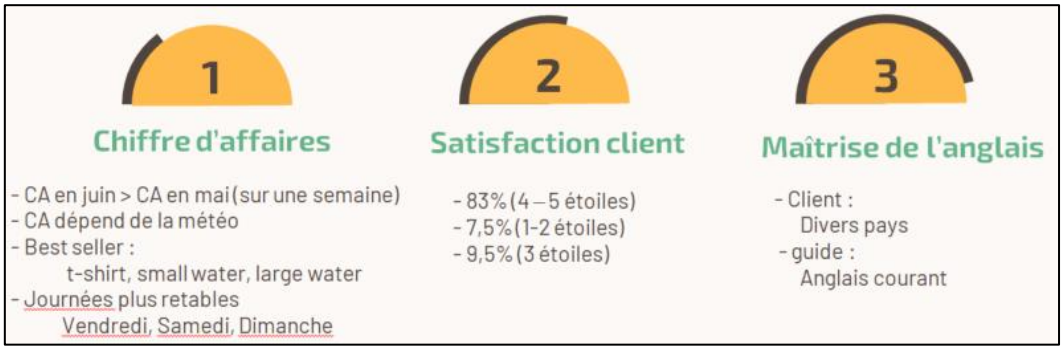


Impact des variations climatiques sur la vente

VI. Conclusion

Impact attendu :

- Amélioration de l'activité (L'influence de la météo sur le chiffre d'affaires)
- Proposition de nouveaux plans d'action (La satisfaction client, maîtrise de l'anglais par les guides)



Axes d'amélioration :