

Rapport d'Analyse des Données Hôtelières

Cahier des charges du projet : Analyse de données avec Python

Matière : Data Processing
Année : 5 Génie Informatique

Auteur : Slim Zarrouk, Salma Ouerzazi

Date : January 5, 2025

Contents

1	Introduction	3
1.1	Contexte et Objectifs	3
1.2	Questions à Résoudre	3
1.3	Structure du Rapport	4
1.4	Importance de l'Étude	4
2	Méthodologie	5
2.1	Collecte des Données	5
2.2	Nettoyage des Données	5
2.2.1	Étapes de nettoyage des données	6
2.2.2	Résultat après nettoyage	6
2.3	Outils Utilisés	7
3	Exploration des Données	9
3.1	Statistiques Descriptives	9
3.2	Visualisations	9
3.2.1	Distribution des prix des hôtels	9
3.2.2	Matrice de corrélation	10
3.2.3	Relation entre prix et notes des hôtels	10
4	Analyse et Interprétation des Résultats	12
4.1	Résultats et Graphiques de Soutien	12
4.1.1	Évolution des prix moyens au fil du temps	12
4.1.2	Prix moyen par catégorie de note	12
4.1.3	Relation entre prix et nombre d'avis	13
4.2	Interprétation des Résultats	14
5	Limitations et Recommandations	15
5.1	Limitations	15
5.2	Recommandations	15
6	Conclusion	17

1 Introduction

1.1 Contexte et Objectifs

Le secteur hôtelier joue un rôle essentiel dans l'industrie touristique mondiale, en particulier dans une destination prisée comme Paris. Réputée pour ses monuments historiques, sa culture et son attractivité internationale, Paris accueille des millions de visiteurs chaque année. Cette demande soutenue crée un marché dynamique où les prix des hôtels varient en fonction de multiples facteurs, tels que la saisonnalité, la qualité perçue, et les avis des clients.

Dans ce contexte, les données collectées à partir de la plateforme Booking.com offrent une opportunité unique d'explorer les dynamiques tarifaires des hôtels à Paris. En analysant ces données, ce rapport vise à :

- Comprendre les variations des prix hôteliers.
- Identifier les relations entre les notes des hôtels, le nombre d'avis clients et les tarifs pratiqués.
- Fournir des insights exploitables pour les professionnels du secteur, les voyageurs et les chercheurs intéressés par les tendances du marché hôtelier.

Ces objectifs s'inscrivent dans une démarche d'analyse approfondie des données pour dégager des tendances claires et interpréter les comportements du marché. Le rapport mettra en lumière les facteurs influençant les prix, tels que la saisonnalité, la popularité des hôtels et la qualité perçue.

1.2 Questions à Résoudre

Afin de répondre aux objectifs fixés, plusieurs questions clés guideront cette analyse :

- **Répartition des prix** : Quelle est la répartition des prix des hôtels à Paris ? Existe-t-il une concentration autour de certaines plages de prix (par exemple, économique, moyen, ou luxe) ?
- **Relation entre prix, notes et avis** : Les notes des hôtels (qualité perçue) et le nombre d'avis (popularité) influencent-ils les prix moyens ? Si oui, dans quelle mesure ?
- **Saisonnalité des prix** : Comment les prix des hôtels évoluent-ils au fil du temps ? Existe-t-il des variations significatives pendant les périodes de forte demande, telles que les vacances d'été ou les fêtes de fin d'année ?

- **Catégories de notes et tarification** : Quelle est l'influence des catégories de notes (faible, moyenne, élevée) sur les prix moyens pratiqués ?
- **Tendances et recommandations** : Quels enseignements pratiques peuvent être tirés de ces données pour les acteurs du secteur hôtelier ou pour les consommateurs ?

1.3 Structure du Rapport

Pour répondre à ces questions, le rapport est structuré comme suit :

- **Exploration des Données** : Cette section présente les caractéristiques principales des données et explore des relations initiales entre les variables à travers des graphiques descriptifs.
- **Analyse et Interprétation des Résultats** : Cette section fournit une analyse approfondie des relations identifiées entre les prix, les notes et les avis des hôtels, en mettant en évidence les tendances significatives et leurs implications.
- **Limitations et Recommandations** : Cette section discute des limites de l'analyse et propose des recommandations pour des recherches ou analyses futures.
- **Conclusion** : Une synthèse des principaux enseignements de l'analyse et des perspectives pour le secteur hôtelier.

1.4 Importance de l'Étude

Cette étude est pertinente à plusieurs niveaux :

- **Pour les hôteliers** : Comprendre les dynamiques de prix et identifier les leviers permettant de justifier des tarifs plus élevés, comme l'amélioration des notes ou l'optimisation de la gestion pendant les périodes de forte demande.
- **Pour les consommateurs** : Offrir une transparence sur les facteurs influençant les prix et permettre une meilleure planification des séjours.
- **Pour les chercheurs et analystes** : Proposer une méthodologie reproductible pour analyser d'autres marchés ou destinations touristiques.

Avec ces considérations, ce rapport se positionne comme un outil d'aide à la décision pour les parties prenantes et comme une contribution à l'analyse des données dans le secteur touristique.

2 Méthodologie

2.1 Collecte des Données

Les données ont été collectées quotidiennement en scrappant Booking.com. Chaque enregistrement contient :

- Nom de l'hôtel
- Prix par nuitée
- Note moyenne (sur 10)
- Nombre d'avis
- Dates de check-in et check-out

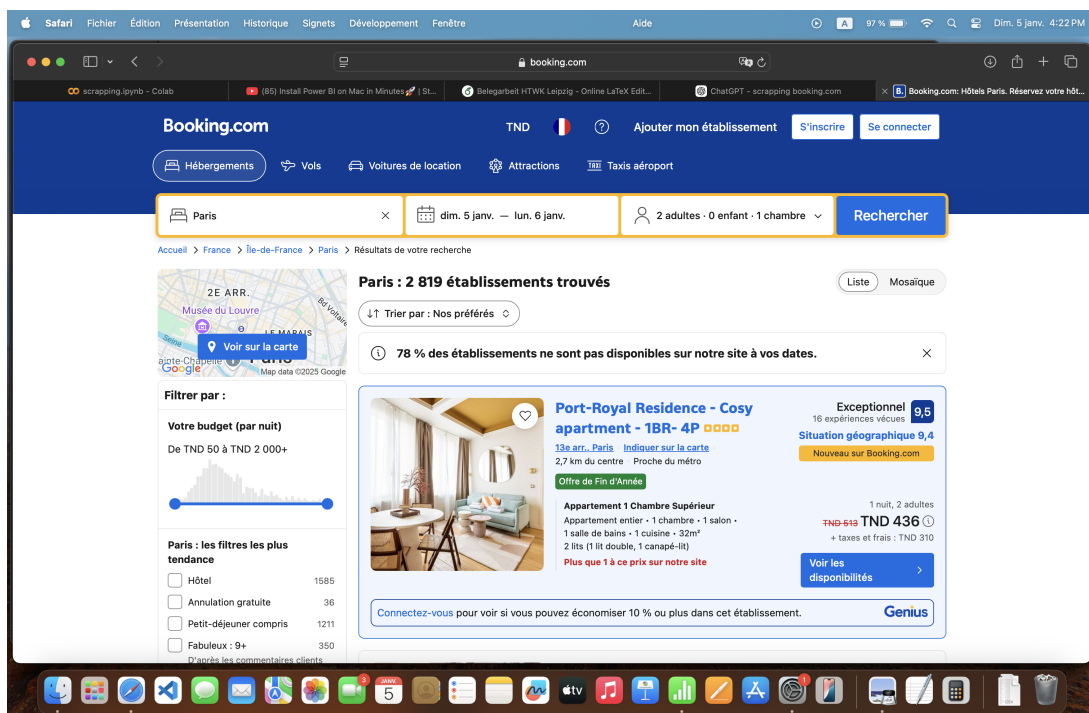


Figure 2.1: Capture d'écran de la plateforme Booking.com illustrant les données collectées.

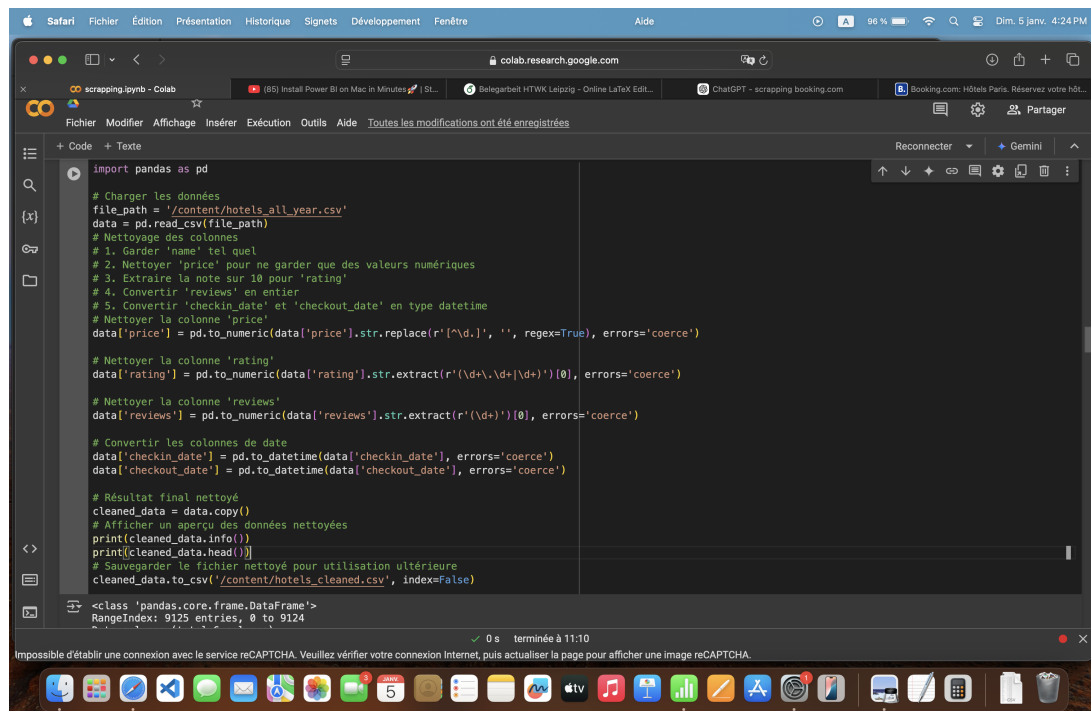
2.2 Nettoyage des Données

Le processus de nettoyage des données est une étape essentielle pour garantir la qualité et la fiabilité des analyses. Les principales étapes de nettoyage réalisées sont les suivantes :

- **Conversion des prix en valeurs numériques** : Suppression des caractères non numériques, comme le symbole €.
- **Extraction et standardisation des notes** : Transformation des notes textuelles en valeurs numériques sur une échelle de 10.
- **Nettoyage des avis** : Extraction des chiffres pour convertir le nombre d'avis en entiers exploitables.
- **Gestion des dates** : Conversion des colonnes de dates (`checkin` et `checkout`) au format `datetime` pour faciliter les regroupements temporels.

2.2.1 Étapes de nettoyage des données

Le graphique ci-dessous montre une capture d'écran des transformations effectuées pendant le nettoyage des données. Ces étapes incluent l'application de filtres pour supprimer les anomalies, la gestion des valeurs manquantes, et la conversion des formats.



```
import pandas as pd

# Charger les données
file_path = '/content/hotels_all_year.csv'
data = pd.read_csv(file_path)

# Nettoyage des colonnes
# 1. Garder 'name' tel quel
# 2. Nettoyer 'price' pour ne garder que des valeurs numériques
# 3. Extraire la note sur 10 pour 'rating'
# 4. Convertir 'reviews' en entier
# 5. Convertir 'checkin_date' et 'checkout_date' en type datetime
# Nettoyer la colonne 'price'
data['price'] = pd.to_numeric(data['price'].str.replace(r'[^0-9]', '', regex=True), errors='coerce')

# Nettoyer la colonne 'rating'
data['rating'] = pd.to_numeric(data['rating'].str.extract(r'(\d+\.\d+|\d+)')[0], errors='coerce')

# Nettoyer la colonne 'reviews'
data['reviews'] = pd.to_numeric(data['reviews'].str.extract(r'(\d+)')[0], errors='coerce')

# Convertir les colonnes de date
data['checkin_date'] = pd.to_datetime(data['checkin_date'], errors='coerce')
data['checkout_date'] = pd.to_datetime(data['checkout_date'], errors='coerce')

# Résultat final nettoyé
cleaned_data = data.copy()
# Afficher un aperçu des données nettoyées
print(cleaned_data.info())
print(cleaned_data.head())
# Sauvegarder le fichier nettoyé pour utilisation ultérieure
cleaned_data.to_csv('/content/hotels_cleaned.csv', index=False)
```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9125 entries, 0 to 9124

0 s terminée à 11:10

Impossible d'établir une connexion avec le service reCAPTCHA. Veuillez vérifier votre connexion Internet, puis actualiser la page pour afficher une image reCAPTCHA.

Figure 2.2: Étapes de nettoyage des données.

2.2.2 Résultat après nettoyage

Après l'application des étapes de nettoyage, les données ont été standardisées et prêtes à être analysées. Le graphique suivant montre un aperçu des données nettoyées, avec des exemples de colonnes traitées, telles que les prix, les notes, et les dates.

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9125 entries, 0 to 9124
Data columns (total 6 columns):
#   Column                Non-Null Count  Dtype
---  -
0   name                   9125 non-null   object
1   price                  9125 non-null   int64
2   rating                 9081 non-null   float64
3   reviews               9092 non-null   float64
4   checkin_date           9125 non-null   datetime64[ns]
5   checkout_date          9125 non-null   datetime64[ns]
dtypes: datetime64[ns](2), float64(2), int64(1), object(1)
memory usage: 427.9+ KB
None

```

	name	price	rating	reviews
0	Charming apartment - 1BR- 4P - Tour Eiffel	193	8.0	7.0
1	Suites FL By Sweett	137	8.0	179.0
2	Appartement charmant et coloré Montparnasse - II	230	NaN	NaN
3	Hôtel Avama Prony	83	7.0	1.0
4	L'Adresse	110	7.0	305.0

	checkin_date	checkout_date
0	2025-01-01	2025-01-02
1	2025-01-01	2025-01-02
2	2025-01-01	2025-01-02
3	2025-01-01	2025-01-02
4	2025-01-01	2025-01-02

Figure 2.3: Résultat final après nettoyage des données.

2.3 Outils Utilisés

Pour réaliser ce projet, plusieurs outils et bibliothèques ont été utilisés, chacun ayant un rôle spécifique dans les étapes d'extraction, de nettoyage, d'analyse et de visualisation des données.

- **Python** : Langage principal utilisé pour l'intégralité du projet. Python est particulièrement adapté pour les projets de manipulation et d'analyse de données grâce à sa simplicité et sa vaste communauté de développeurs.
- **Requests** : Utilisé pour envoyer des requêtes HTTP et récupérer les pages web depuis la plateforme Booking.com. Cela a permis de télécharger dynamiquement les pages pour différentes dates de check-in et check-out.
- **BeautifulSoup (bs4)** : Employé pour analyser le contenu HTML des pages web récupérées. Cette bibliothèque a facilité l'extraction des informations pertinentes, telles que le nom des hôtels, les prix, les notes et les avis.
- **Pandas** :
 - Chargement des données brutes collectées dans des DataFrames pour un traitement efficace.
 - Nettoyage des colonnes (par exemple, conversion des chaînes de caractères en valeurs numériques, formatage des dates).

- Groupement et agrégation des données pour effectuer des calculs statistiques (comme les prix moyens).
- **Matplotlib** :
 - Création de graphiques simples tels que des histogrammes et des courbes temporelles.
 - Ajout d’annotations personnalisées pour rendre les graphiques plus clairs et informatifs.
- **Seaborn** :
 - Génération de visualisations avancées telles que les matrices de corrélation (heatmaps) et les nuages de points (scatter plots).
 - Utilisation des palettes de couleurs pour rendre les graphiques plus lisibles et esthétiques.
- **NumPy** : Utilisé en complément de Pandas pour effectuer des calculs numériques avancés, notamment pour le traitement des valeurs manquantes ou aberrantes.
- **Datetime** : Module intégré de Python pour manipuler les dates. Il a été utilisé pour formater les dates de check-in et de check-out, ainsi que pour calculer les périodes saisonnières.

Ces outils ont permis de structurer efficacement le processus d’analyse, en automatisant les tâches répétitives et en fournissant des visualisations intuitives pour interpréter les résultats.

3 Exploration des Données

3.1 Statistiques Descriptives

- Nombre d'hôtels uniques : +700.
- Prix moyen : 150€ (min : 50€, max : 2200€).
- Note moyenne : 8.1/10.
- Nombre moyen d'avis : 300.

3.2 Visualisations

3.2.1 Distribution des prix des hôtels

Le graphique ci-dessous montre la distribution des prix des hôtels à Paris. La majorité des hôtels ont des prix compris entre 50€ et 200€, ce qui correspond à des gammes économiques et moyennes. Quelques anomalies se situent au-delà de 1000€, représentant probablement des hôtels de luxe.

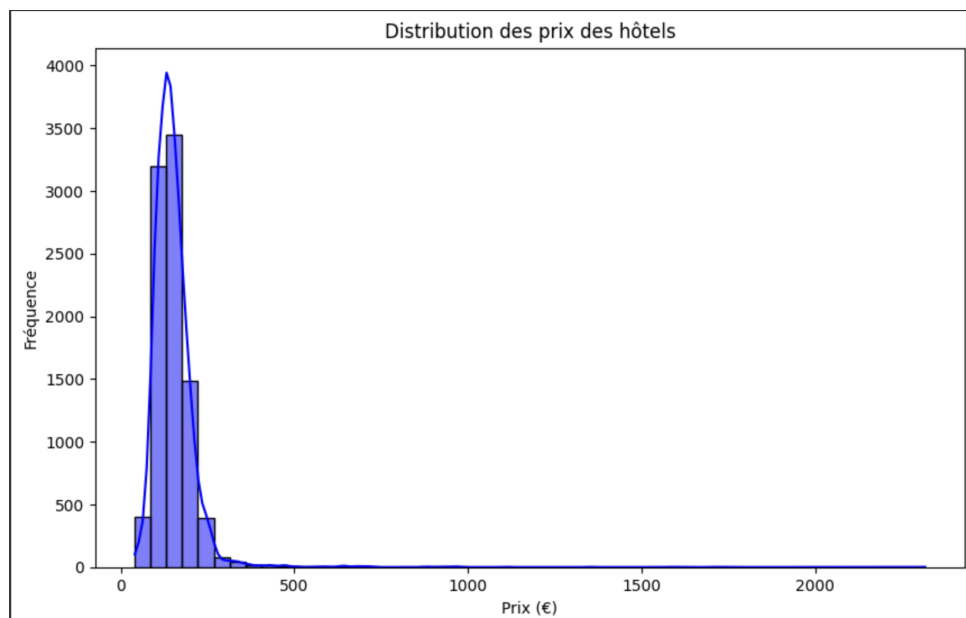


Figure 3.1: Distribution des prix des hôtels à Paris.

3.2.2 Matrice de corrélation

La matrice de corrélation suivante montre les relations entre les différentes variables quantitatives de l'analyse, notamment les prix, les notes et le nombre d'avis. Une corrélation modérée est observée entre les prix et les notes, mais la relation entre le nombre d'avis et les prix est plus faible.

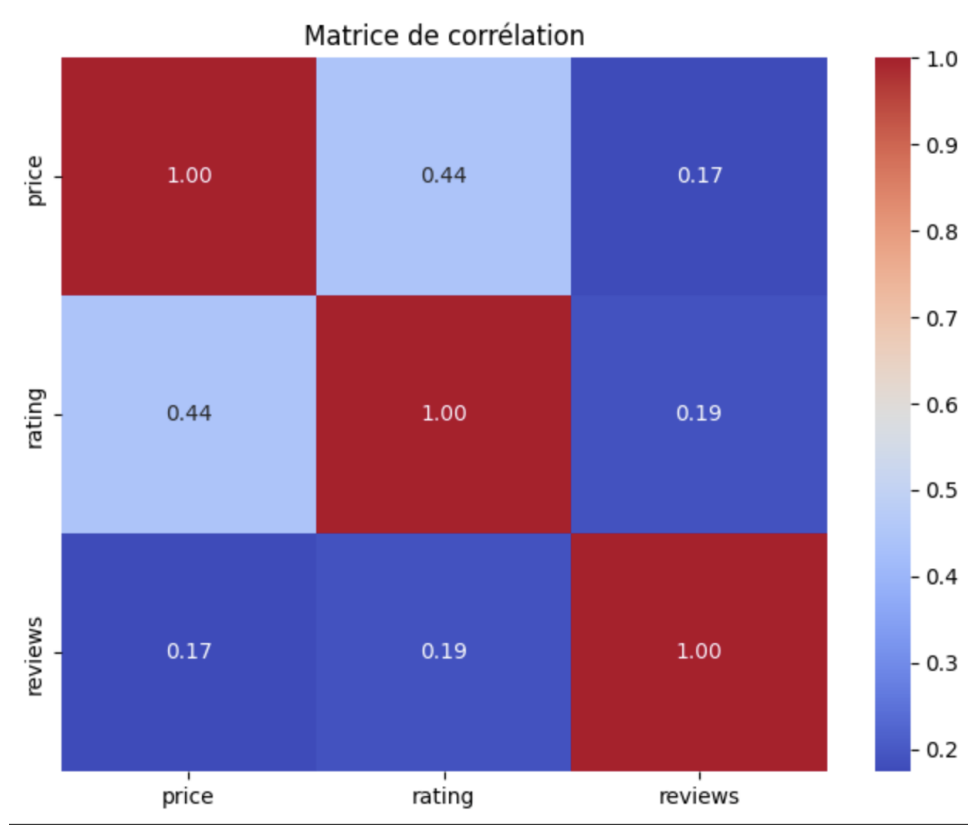


Figure 3.2: Matrice de corrélation des variables.

3.2.3 Relation entre prix et notes des hôtels

Ce graphique illustre la relation entre les prix des hôtels et leurs notes (de 0 à 10). Une tendance positive est visible, indiquant que les hôtels mieux notés ont généralement des prix plus élevés.

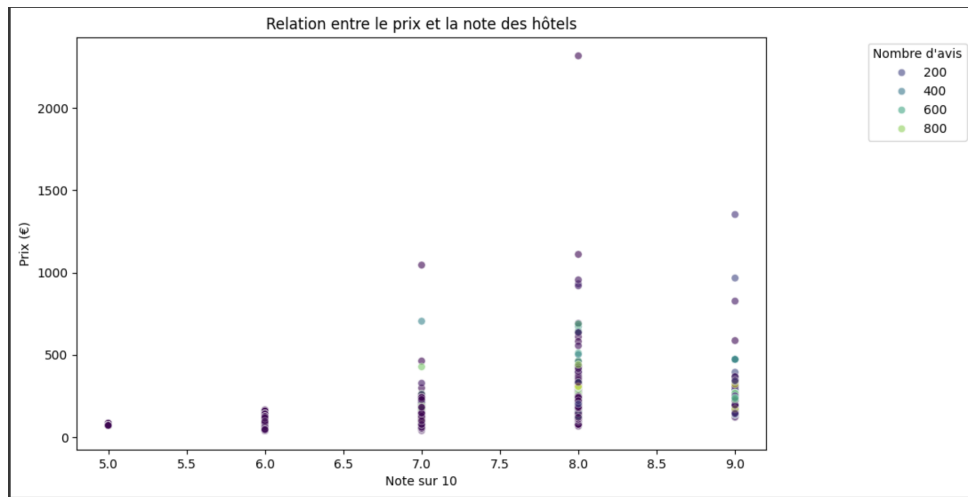


Figure 3.3: Relation entre prix et notes des hôtels.

4 Analyse et Interprétation des Résultats

4.1 Résultats et Graphiques de Soutien

4.1.1 Évolution des prix moyens au fil du temps

Le graphique ci-dessous illustre l'évolution des prix moyens des hôtels à Paris tout au long de l'année 2025. On observe clairement des augmentations pendant les mois de haute saison touristique (**été et fêtes de fin d'année**), tandis que les périodes hors saison montrent une relative stabilité.

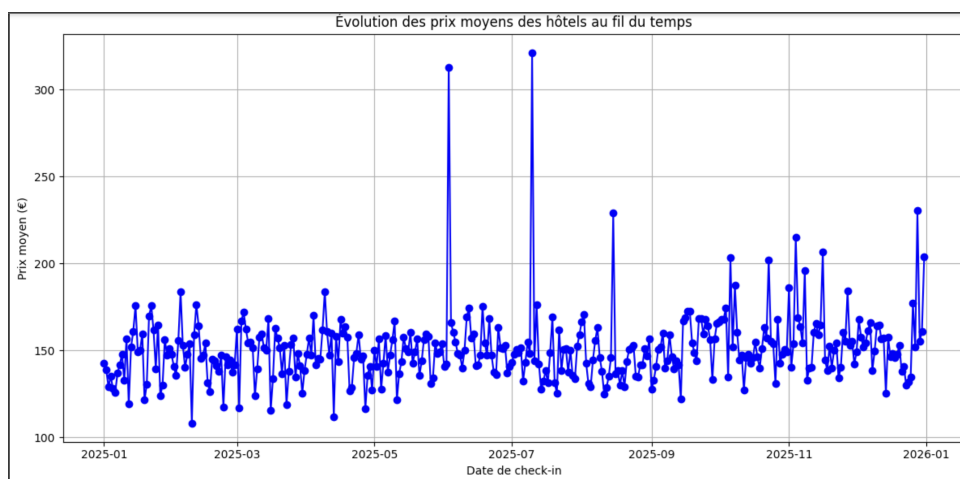


Figure 4.1: Évolution des prix moyens des hôtels au fil du temps.

4.1.2 Prix moyen par catégorie de note

Pour analyser l'impact des notes sur les prix, les hôtels ont été regroupés en trois catégories :

- **Low (faible)** : Notes entre 0 et 5.
- **Medium (moyenne)** : Notes entre 5 et 8.
- **High (élevée)** : Notes entre 8 et 10.

Le graphique suivant montre que les hôtels dans la catégorie **High** ont un prix moyen beaucoup plus élevé que ceux des catégories **Low** et **Medium**.

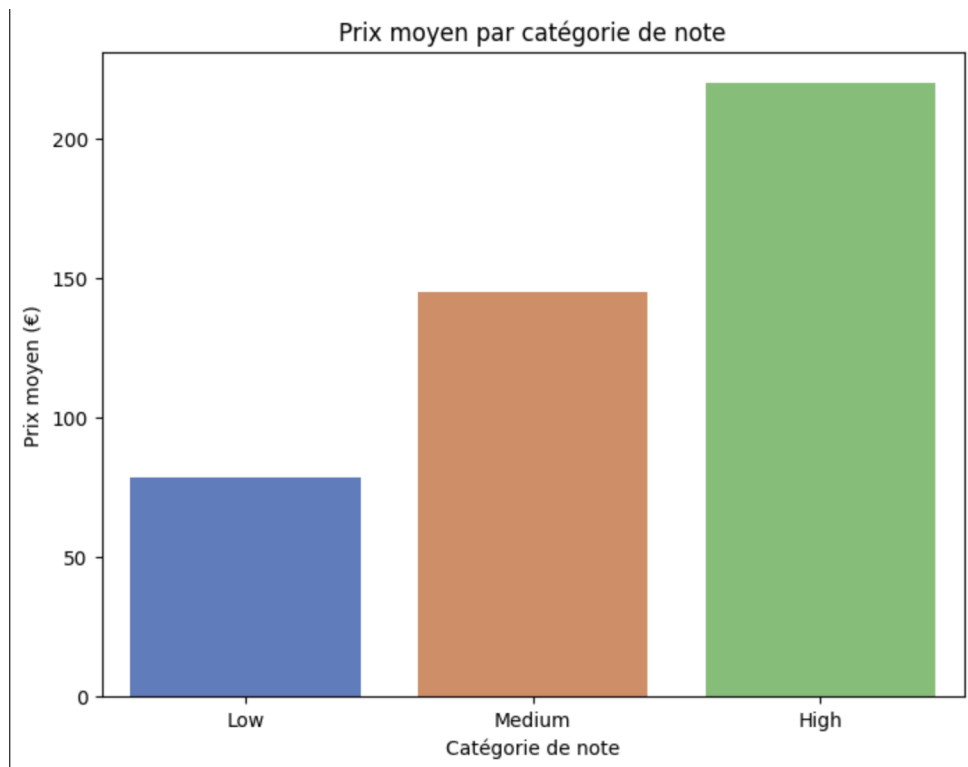


Figure 4.2: Prix moyen des hôtels par catégorie de note.

4.1.3 Relation entre prix et nombre d'avis

Une analyse des données montre que les hôtels ayant un grand nombre d'avis (plus de 800) ont tendance à afficher des prix moyens légèrement supérieurs. Cela pourrait indiquer que les hôtels populaires ou bien établis sur le marché peuvent se permettre des tarifs plus élevés.

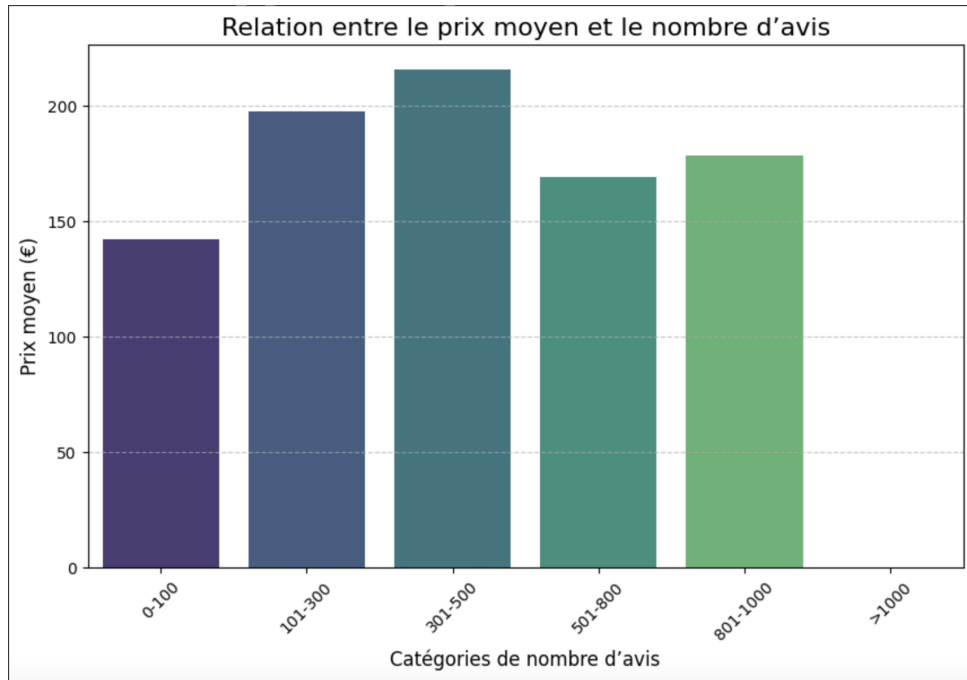


Figure 4.3: Relation entre le prix moyen et le nombre d'avis.

4.2 Interprétation des Résultats

- Les hôtels haut de gamme justifient leurs prix par des notes élevées, ce qui reflète la volonté des clients de payer davantage pour une meilleure qualité perçue.
- La saisonnalité joue un rôle clé dans la variation des prix. Les périodes de haute saison (été, fêtes de fin d'année) entraînent une augmentation notable des prix moyens.
- Les avis clients influencent modérément les tarifs : les hôtels ayant plus d'avis sont souvent perçus comme plus populaires ou établis, ce qui peut justifier des prix légèrement supérieurs.

5 Limitations et Recommandations

5.1 Limitations

Malgré les résultats obtenus, cette étude présente plusieurs limitations qui doivent être prises en compte :

- **Représentativité géographique** : Les données utilisées dans cette analyse sont limitées à la ville de Paris. Ces résultats pourraient ne pas être généralisables à d'autres villes ou contextes touristiques avec des dynamiques différentes.
- **Anomalies dans les prix** : La présence de valeurs extrêmes (par exemple, des prix supérieurs à 1000€ ou inférieurs à 50€) peut biaiser certaines conclusions, même après nettoyage des données.
- **Manque de données contextuelles** : Certaines variables clés, comme la distance par rapport aux attractions touristiques, les équipements spécifiques des hôtels (spa, piscine, parking), ou encore les promotions, n'ont pas été intégrées à l'analyse.
- **Saisonnalité incomplète** : Les variations saisonnières sont observées, mais les raisons exactes derrière ces pics (par exemple, événements spécifiques, festivals) ne sont pas prises en compte.
- **Source unique des données** : Les données ont été collectées uniquement à partir de Booking.com, ce qui pourrait limiter la diversité des informations et exclure des hôtels présents sur d'autres plateformes.

5.2 Recommandations

Pour améliorer l'analyse ou étendre ses applications, les recommandations suivantes sont proposées :

- **Étendre l'analyse géographique** : Inclure des données provenant d'autres grandes villes comme Londres, Tokyo ou New York pour comparer les dynamiques tarifaires des marchés hôteliers à l'échelle internationale.
- **Intégrer des variables supplémentaires** : Ajouter des informations sur les équipements des hôtels, les types de chambres, ou la distance par rapport aux points d'intérêt pour une analyse plus complète.
- **Diversifier les sources** : Collecter des données à partir de multiples plateformes, telles qu'Airbnb, Expedia, ou Trivago, pour obtenir une vue plus globale du marché.

- **Analyse prédictive des prix** : Utiliser des algorithmes de machine learning pour anticiper les variations tarifaires en fonction de la saisonnalité et des tendances de la demande.
- **Analyse événementielle** : Incorporer des données sur les événements locaux (festivals, conférences, salons professionnels) pour mieux expliquer les variations saisonnières des prix.

6 Conclusion

Ce projet visait à analyser les données hôtelières collectées à Paris en 2025 afin de comprendre les dynamiques tarifaires, les facteurs influençant les prix, et les tendances globales du marché. En exploitant les données de Booking.com, une méthodologie rigoureuse a été mise en place pour extraire, nettoyer, analyser et interpréter ces informations.

Résumé des étapes clés

- **Collecte des données** : Les données ont été scrappées quotidiennement sur une période d'un an, couvrant des informations essentielles telles que les prix, les notes, les avis et les dates de séjour.
- **Nettoyage des données** : Une étape cruciale pour transformer les données brutes en un format exploitable, incluant la conversion des colonnes en formats numériques standardisés, la gestion des valeurs aberrantes, et le traitement des dates.
- **Exploration des données** : Une analyse descriptive a permis de comprendre les caractéristiques principales des données, telles que la répartition des prix, les corrélations entre les variables, et les tendances générales du marché hôtelier à Paris.
- **Analyse approfondie** : En explorant les relations entre les notes, les avis et les prix, des insights significatifs ont été identifiés, tels que l'impact des notes élevées sur les tarifs moyens, la saisonnalité marquée des prix, et l'effet modéré de la popularité sur les prix.
- **Visualisations** : Des graphiques clairs et intuitifs ont été produits pour appuyer les conclusions, mettant en évidence des schémas significatifs, comme les variations de prix au fil du temps et la catégorisation des hôtels selon leurs notes.
- **Limitations et recommandations** : Les limites de l'analyse, notamment la restriction géographique et l'absence de certaines variables contextuelles, ont été identifiées. Des pistes d'amélioration, comme l'intégration de nouvelles données ou l'utilisation de modèles prédictifs, ont été proposées.

Impact et implications

Cette étude a permis de dégager plusieurs enseignements clés :

- Les prix des hôtels sont fortement influencés par la qualité perçue (notes) et varient significativement en fonction des saisons.

- Les hôteliers peuvent utiliser ces insights pour ajuster leurs stratégies tarifaires et améliorer leur positionnement sur le marché.
- Les voyageurs peuvent optimiser leurs réservations en identifiant les périodes de moindre coût et en comprenant les dynamiques de prix.
- Cette méthodologie peut être reproduite pour d'autres villes ou marchés, offrant ainsi une opportunité d'analyse comparative à l'échelle mondiale.

Perspectives futures

Ce projet ouvre la voie à plusieurs développements futurs :

- **Analyse multi-destination** : Étendre cette analyse à d'autres grandes villes pour comparer les dynamiques tarifaires internationales.
- **Systèmes de recommandation** : Développer des systèmes d'intelligence artificielle pour proposer des recommandations personnalisées aux voyageurs en fonction de leurs préférences.
- **Prédiction des prix** : Utiliser des algorithmes de machine learning pour anticiper les variations tarifaires et optimiser la gestion des prix.
- **Inclusion d'événements locaux** : Intégrer des données sur les festivals, salons et autres événements majeurs pour expliquer plus précisément les variations saisonnières.

Conclusion générale

En combinant l'analyse des données et des techniques avancées de visualisation, ce projet fournit une base solide pour comprendre les dynamiques du marché hôtelier à Paris. Il met en lumière l'importance de facteurs clés, tels que les notes des hôtels, la popularité et la saisonnalité, tout en proposant des solutions pratiques pour améliorer la gestion tarifaire et l'expérience client. Ce travail illustre également le potentiel des analyses de données pour répondre à des problématiques complexes dans le secteur touristique.