Introduction

Contexte et Objectifs :

Avec la montée en popularité des comparateurs de vols en ligne, accéder à des informations claires et pertinentes est crucial pour les voyageurs cherchant à optimiser leurs dépenses. L'objectif principal de cette analyse est de développer un outil automatisé permettant de collecter des données sur les vols et d'offrir aux utilisateurs des informations détaillées, comme les prix moyens, les durées et les types de vols, pour prendre des décisions éclairées.

Problématique :

Comment concevoir un outil permettant aux utilisateurs de trouver des prix de vols plus abordables tout en assurant la prise en compte d'autres critères essentiels, tels que la durée des vols, les types de vols, la disponibilité, et la fluctuation des prix en fonction des différentes caractéristiques comme la destination, le pays d'origine, et d'autres facteurs variables, tout en garantissant la collecte, le nettoyage et l'analyse des données de manière fiable et précise ?

Cette problématique permet une analyse structurée et offre une approche claire pour examiner non seulement les prix des vols, mais aussi les variables qui influencent ces prix, telles que la durée des vols, les types de vols (directs, avec escale), ainsi que la saisonnalité et la disponibilité des vols. Elle inclut également des aspects comme le pays d'origine (Tunisie dans ce cas) et la destination, ce qui permet de donner une vision complète des facteurs déterminants dans le calcul du prix d'un vol. Cela laisse place à une exploration détaillée des données récupérées, avec une analyse statistique et la visualisation de ces dernières pour en tirer des conclusions pertinentes.

Introduction de Kayak :

Kayak est un site de comparaison de prix de voyages qui collecte et analyse les informations sur les billets d'avion, hôtels, et locations de voitures à travers plusieurs plateformes. Il propose une interface relativement simple en termes de

code HTML et présente moins de complexités dynamiques par rapport à d'autres sites comme Google Flights. Cela rend Kayak un choix approprié pour des projets de scraping, notamment lorsqu'il s'agit d'extraire des informations de manière automatisée.

Choix de Kayak et Raisons :

Nous avons opté pour Kayak pour le projet de scraping des prix de vols et des informations associées principalement en raison de la structure HTML relativement simple de son site, ainsi que de l'utilisation modérée de JavaScript et d'éléments dynamiques. Ce site est plus facile à analyser par rapport à d'autres options comme Google Flights, qui possède une structure HTML complexe et nécessite une gestion beaucoup plus élaborée des éléments dynamiques. Une autre option, Tunisie Booking, s'est avérée impraticable à cause de l'URL générée de manière aléatoire, rendant difficile la réutilisation du même lien pour les tests.

Choix des Outils et Techniques :

Pour effectuer le scraping, nous avons utilisé **Selenium**, un outil qui permet de contrôler un navigateur web de manière automatique. Selenium est particulièrement adapté aux sites avec des éléments dynamiques, comme ceux de Kayak, qui utilisent des boutons et des composants interactifs, ainsi que des informations dynamiques. De plus, Kayak met en place des mesures anti-bots assez élaborées, telles que des Captchas et des restrictions basées sur des comportements suspects, rendant l'utilisation de bibliothèques comme **BeautifulSoup** ou **Scrapy** inappropriée pour ce projet. nous avons essayé **bs4 et Scrapy**, mais kayak a des mesures anti bots qui ont détecté ces modules.

En revanche, Selenium, grâce à son mode "headless" (sans interface graphique), a permis de contourner ces protections et de récupérer les informations nécessaires sans trop de difficulté.

Données Scrappées :

L’objectif principal du projet est de créer un outil permettant à l’utilisateur de trouver des prix de vols moins chers tout en maintenant la prise en compte d’autres informations cruciales. Pour ce faire, nous avons concentré notre extraction de données sur les vols au départ de la **Tunisie** à destination d’une **destination fixe** pour limiter la durée d’exécution du programme. Toutefois, le script peut facilement être modifié pour changer la destination ou le pays d'origine.

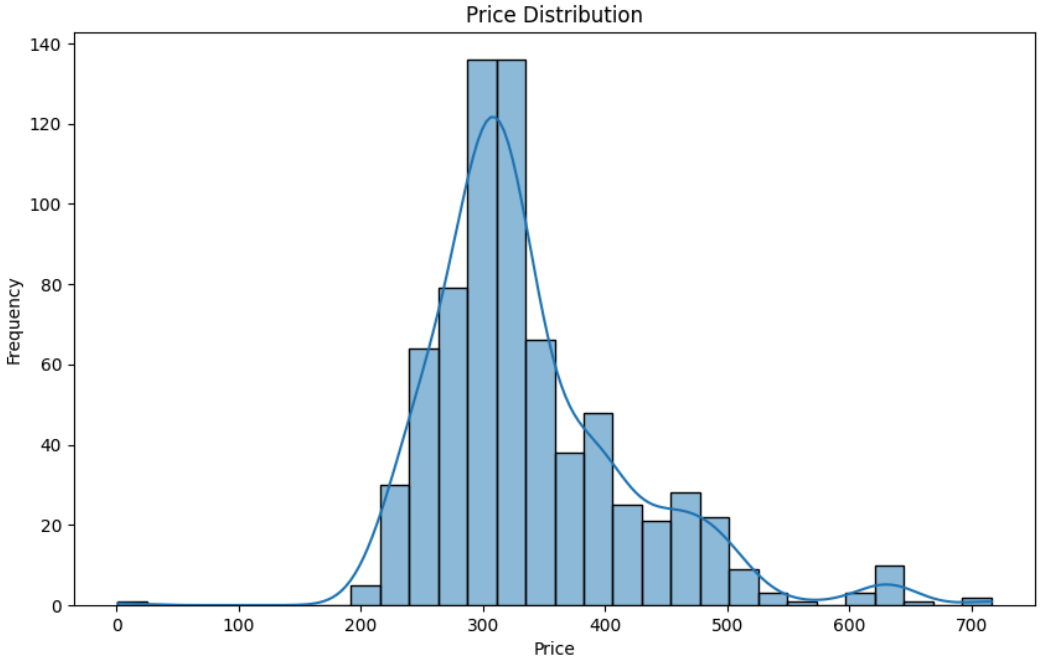
Le script a été exécuté en boucle sur une liste de tous les aéroports tunisiens, collectant ainsi un grand nombre de données. Au cours de cette première phase, nous avons obtenu des milliers de résultats, que nous avons ensuite nettoyés (suppression des doublons, gestion des données manquantes, etc.) pour obtenir un jeu de données cohérent de moins de **mille vols**.

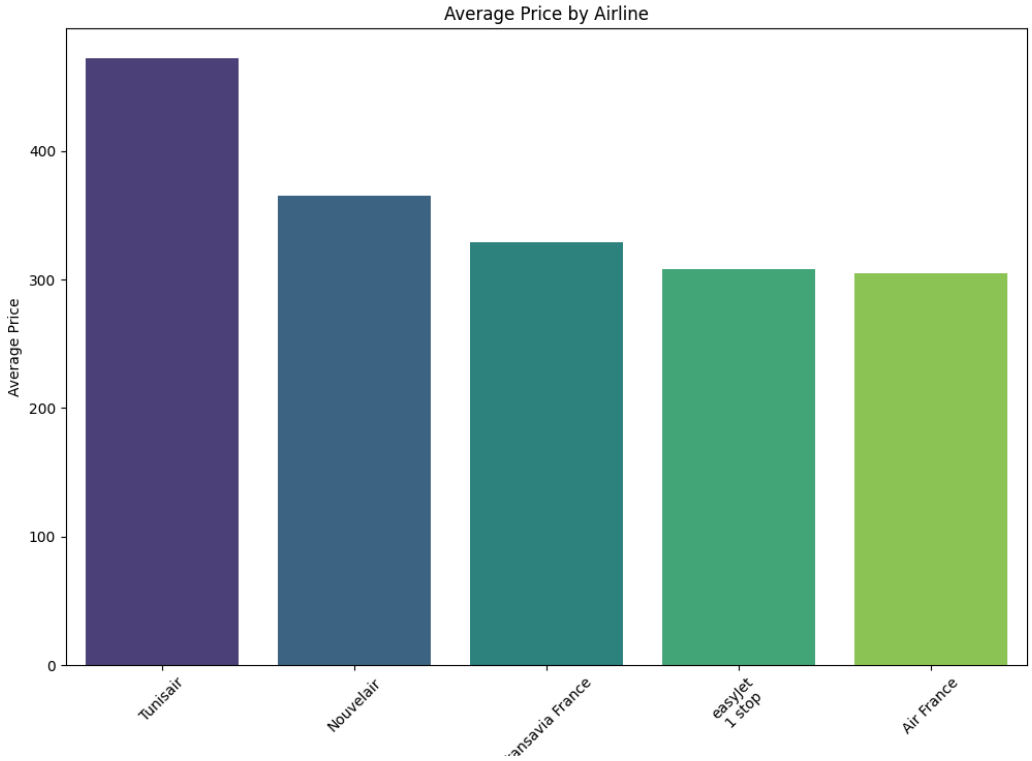
Nettoyage et Préparation des Données :

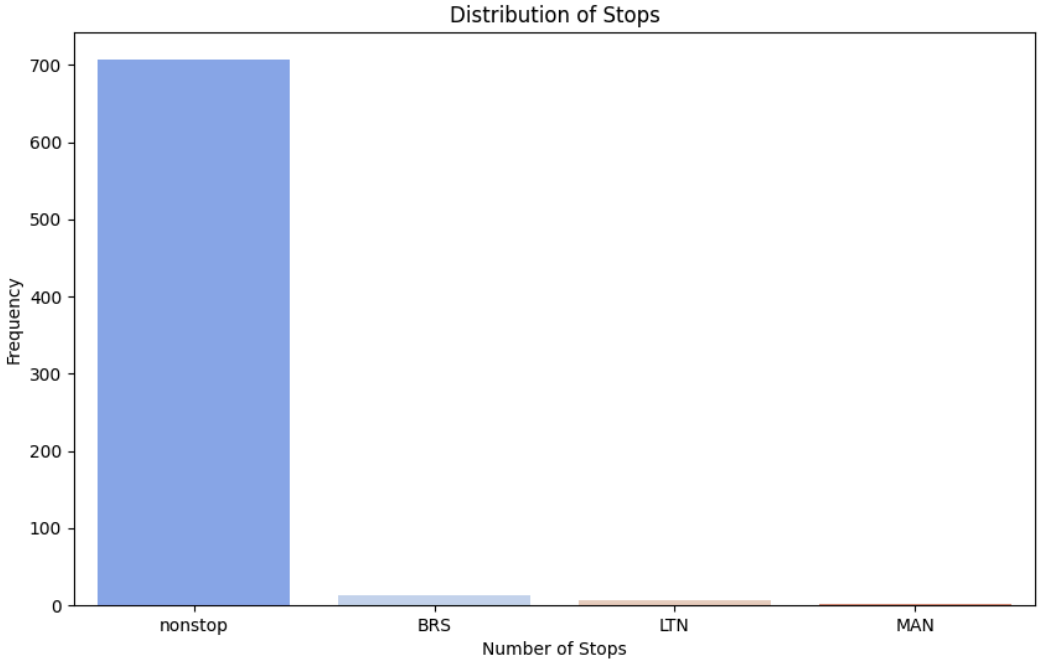
Pour organiser et structurer les données, nous avons utilisé **Pandas** et des **DataFrames**, stockant les informations dans des fichiers **XLSX** avant et après le nettoyage. Le nettoyage des données a consisté à retirer les doublons, gérer les valeurs manquantes et corriger les erreurs, ce qui a permis d’obtenir un jeu de données fiable pour l’analyse.

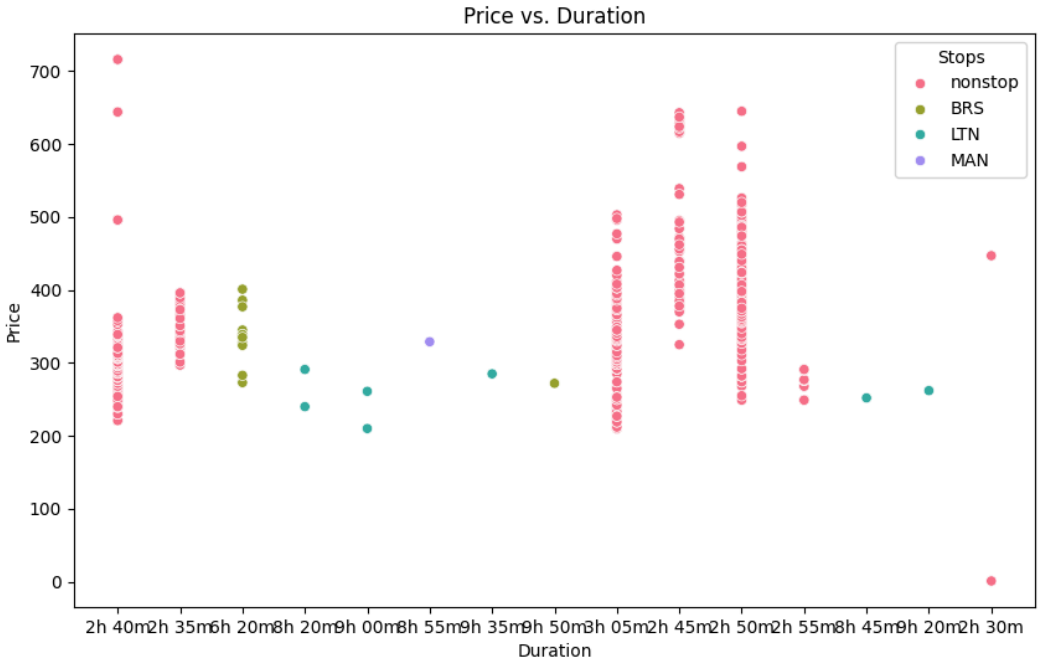
Visualisation et Analyse :

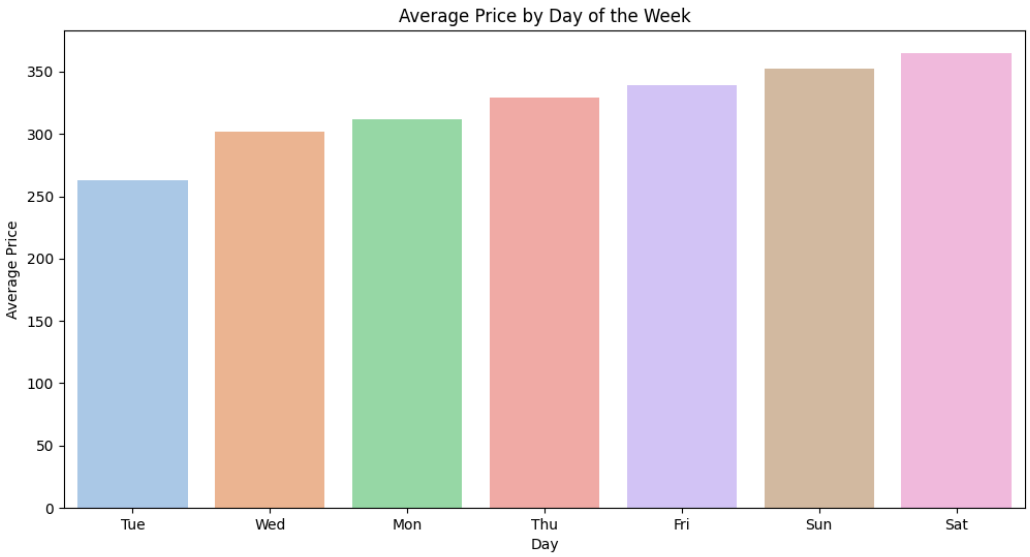
L’analyse des données a été réalisée à l’aide des bibliothèques **Matplotlib** et **Seaborn**. Ces outils ont permis de produire des visualisations telles que des histogrammes, des graphiques de dispersion et des cartes de chaleur pour observer les tendances et anomalies dans les données, notamment en ce qui concerne la répartition des prix, les durées de vol et la relation entre les prix et les autres variables.

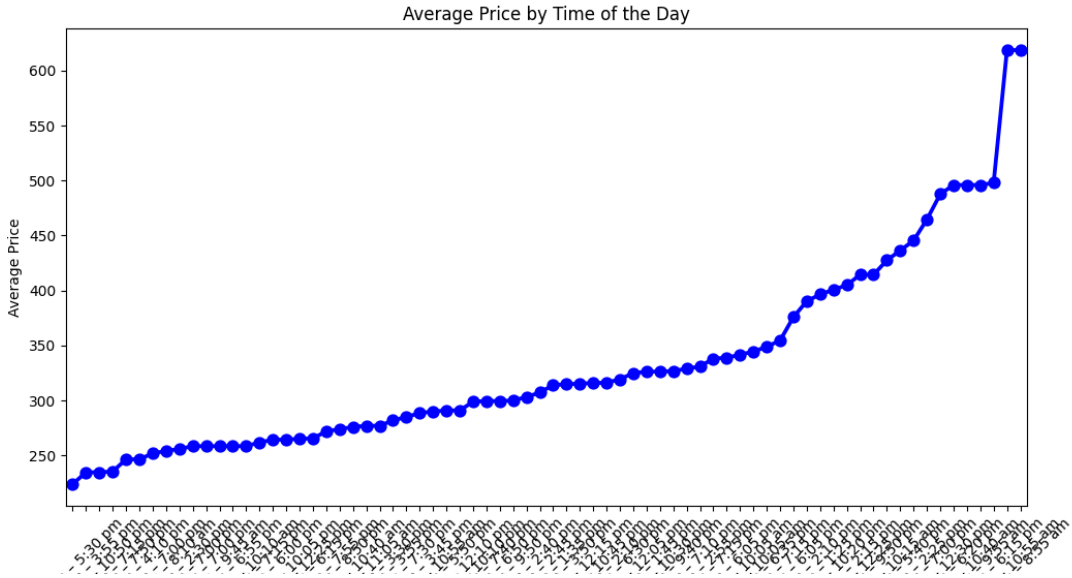

Price Distribution
Frequency
Price
0
20
40
60
80
100
120
140
0
100
200
300
400
500
600
700

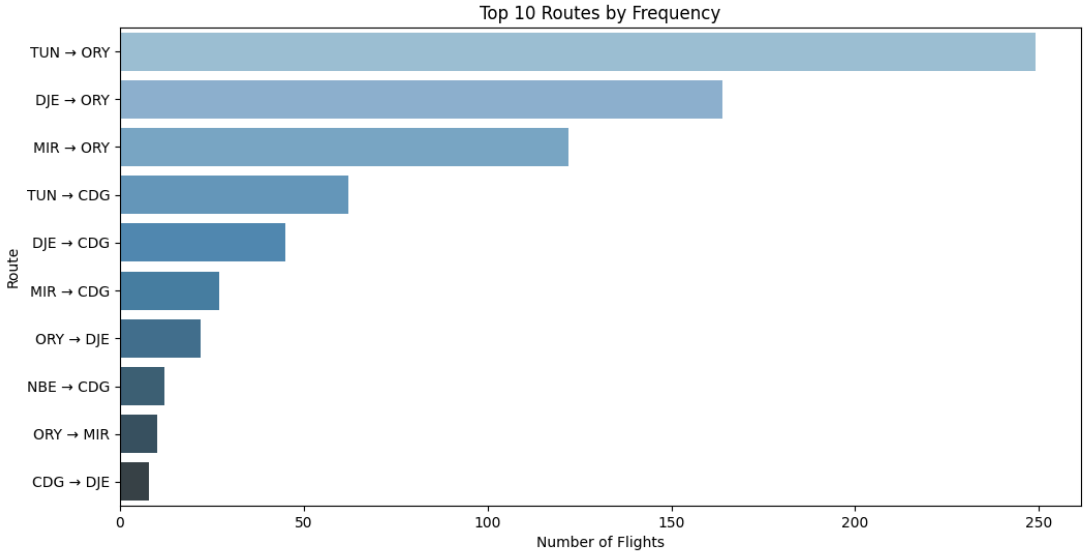

Average Price by Airline
Average Price
0
100
200
300
400
Tunisair
Nouvelair
easylet
1 stop
Air France


Distribution of Stops
Frequency
0
100
200
300
400
500
600
700
nonstop
BRS
LTN
MAN
Number of Stops


Price vs. Duration
Price
0
100
200
300
400
500
600
700
Stops
nonstop
BRS
LTN
MAN
2h 40m
2h 35m
6h 20m
8h 20m
9h 00m
8h 55m
9h 35m
9h 50m
3h 05m
2h 45m
2h 50m
2h 55m
8h 45m
9h 20m
2h 30m
Duration


Average Price by Day of the Week
Average Price
0
50
100
150
200
250
300
350
Tue
Wed
Mon
Thu
Fri
Sun
Sat
Day





Conclusion :

Grâce à cette approche, l'outil permet désormais d'accéder à des informations complètes sur les **prix moyens des vols**, les **types de vols** (directs ou avec escale), ainsi que leur **durée**, tout en tenant compte des fluctuations saisonnières et autres critères pertinents. Le projet a ainsi permis de répondre à la problématique en offrant un moyen de trouver des prix de vols plus abordables tout en conservant des informations fiables et utiles pour l'utilisateur.