Faculté des Sciences et Ingénierie - Sorbonne université

Master Informatique parcours - DAC

Datascience, Learning and Applications

Rapport technique

Etude de données aéronautique

Réalisé par :
Melissa Dahlia Attabi
Abdelkrim Halimi

Mai 2024

Table des matières

Introduction

Dans ce rapport, nous présentons une analyse détaillée des données aéronautiques. Notre étude se base sur des données provenant de Flight Radar, couvrant divers aspects comme les types de compagnies aériennes, les statistiques de vol, les types d'avions utilisés, les destinations desservies et les retards enregistrés. En complément, nous avons également exploré les avis clients pour obtenir des perspectives qualitatives sur les performances des compagnies aériennes et des aéroports. L'objectif principal de ce rapport est de fournir des perspectives claires et exploitables qui peuvent aider à la prise de décision stratégique, notamment dans le contexte de l'ouverture d'une nouvelle compagnie aérienne aux États-Unis, et ce, grâce au développement de modèles de Machine Learning.

Chapitre 1

Acquisition et pré-traitement des données

Dans cette partie du rapport, nous détaillons les techniques utilisées afin de récupérer les différentes données utilisées ainsi que le pré-traitement effectué.

1.1 Description des données

Les données étudiées dans ce projet sont des données de type aéronautique. Elles proviennent d'un site de suivi de vols en temps réel : Flight Radar [**flight**]. Ce service nous permet d'obtenir des informations concernant :

- Aéroports
  - Coordonnées géographiques
  - Avis clients
  - Vols à destination
  - Vols au départ
- Compagnies aériennes
  - Flotte aérienne
- Vols
  - Heure de décollage
  - Heure d'arrivée
  - Compagnie aérienne
  - Aéroport d'arrivée
  - Aéroport de départ
  - Type d'avion

1.2 Acquisition des données

On tente donc de récupérer toutes ces informations du site Flight Radar [**flight**]. On adopte trois méthodes :

1. **Web Scraping avec notre code**
   Nous avons implémenter quatres classes :
   - **AirportScrapper**
     Crée un dataframe python qui contient deux colonnes : le **pays** et l'**aéroport**.

— **DepArrScrapper**
Retourne un dataframe python contenant des informations sur les départs et arrivées des vols :

- Pays
- Aéroport
- Date
- Heure
- Identifiant du vol
- Compagnie aérienne
- Type d'avion
- Immatriculation de l'avion
- Status (Atteri/Prévu/Décalé/...)

— **ReviewScrapper**
Récupère les avis clients (notes sur 5 et commentaires) sur tous les aéroports (lorsqu'ils existent). Le dataframe contient :

- Pays
- Date de l'avis
- Commentaire des clients
- Note moyenne
- Notes pour les différentes catégories (par exemple WiFi, Sécurité, Récupération des bagages,...)

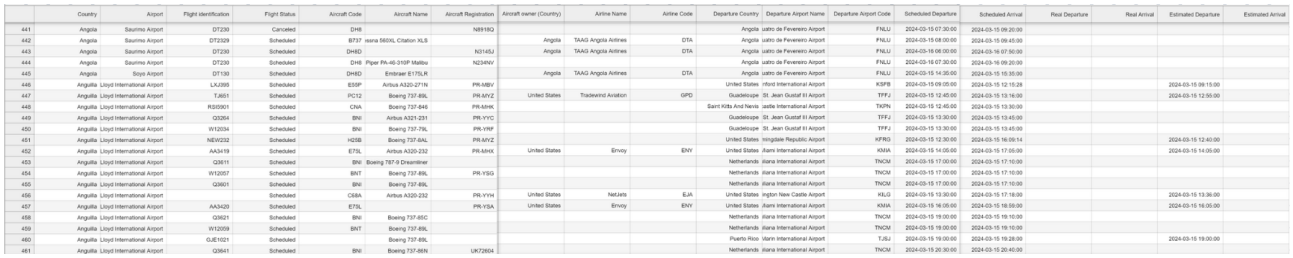
Nous avons opté pour l'utilisation des bibliothèques **request** et **selenium**, accompagnées de **undetected_chromedriver**, étant donnée du caractère dynamique des données de la page, et afin de surmonter les vérifications du site visant à détecter les robots.
Toutefois, comme le temps d'exécution était trop important, nous n'avons donc pas pu tout récupérer qu'avec ce code. Nous avons donc eu recours à une API.

2. **Web Scraping en utilisant une API**
L'API utilisée : FlightRadar24API [**api**] retourne des fichiers .json contenant les mêmes informations que celles cités précédemment.

```
dict_keys(['details', 'flightdiary', 'schedule', 'weather', 'aircraftCount', 'runways'])
dict_keys(['name', 'code', 'delayIndex', 'stats', 'position', 'timezone', 'url', 'visible'])
dict_keys(['url', 'ratings', 'comment', 'reviews', 'evaluation'])
dict_keys(['arrivals', 'departures', 'ground'])
dict_keys(['metar', 'time', 'qnh', 'dewpoint', 'humidity', 'pressure', 'sky', 'flight', 'wind', 'temp', 'elevation', 'cached'])
dict_keys(['ground', 'onGround'])
```

FIGURE 1.1 – Exemple de Dataframe obtenu (Arrivals)

On obtient les dataframes suivants :



	Country	Airport	Flight identification	Flight Status	Aircraft Code	Aircraft Name	Aircraft Registration	Aircraft owner (Country)	Airline Name	Airline Code	Departure Country	Departure Airport Name	Departure Airport Code	Scheduled Departure	Scheduled Arrival	Real Departure	Real Arrival	Estimated Departure	Estimated Arrival
441	Angola	Saurimo Airport	DT230	Canceled	DH8		N818Q				Angola	uatro de Fevereiro Airport	FNLU	2024-03-15 07:30:00	2024-03-15 09:20:00				
442	Angola	Saurimo Airport	DT2329	Scheduled	B737	ssna 560XL Citation XLS		Angola	TAAG Angola Airlines	DTA	Angola	uatro de Fevereiro Airport	FNLU	2024-03-15 08:00:00	2024-03-15 09:45:00				
443	Angola	Saurimo Airport	DT230	Scheduled	DH8D		N314SJ	Angola	TAAG Angola Airlines	DTA	Angola	uatro de Fevereiro Airport	FNLU	2024-03-16 06:00:00	2024-03-16 07:50:00				
444	Angola	Saurimo Airport	DT230	Scheduled	DH8	Piper PA-46-310P Malibu	N234NV				Angola	uatro de Fevereiro Airport	FNLU	2024-03-16 07:30:00	2024-03-16 09:20:00				
445	Angola	Soyo Airport	DT130	Scheduled	DH8D	Embraer E175LR		Angola	TAAG Angola Airlines	DTA	Angola	uatro de Fevereiro Airport	FNLU	2024-03-15 14:35:00	2024-03-15 15:35:00				
446	Anguilla	Lloyd International Airport	LKJ395	Scheduled	ES5P	Airbus A320-271N	PR-MBV				United States	nford International Airport	KSFB	2024-03-15 09:05:00	2024-03-15 12:15:28			2024-03-15 09:15:00	
447	Anguilla	Lloyd International Airport	TJ651	Scheduled	PC12	Boeing 737-89L	PR-MYZ	United States	Tradewind Aviation	GPD	Guadeloupe	St. Jean Gustaf III Airport	TFFJ	2024-03-15 12:45:00	2024-03-15 13:16:00			2024-03-15 12:55:00	
448	Anguilla	Lloyd International Airport	RSI6901	Scheduled	CNA	Boeing 737-846	PR-MHK				Saint Kitts And Nevis	aseterre International Airport	TKPN	2024-03-15 12:45:00	2024-03-15 13:30:00				
449	Anguilla	Lloyd International Airport	Q3264	Scheduled	BNI	Airbus A321-231	PR-YYC				Guadeloupe	St. Jean Gustaf III Airport	TFFJ	2024-03-15 13:30:00	2024-03-15 13:45:00				
450	Anguilla	Lloyd International Airport	W12034	Scheduled	BNI	Boeing 737-79L	PR-YNF				Guadeloupe	St. Jean Gustaf III Airport	TFFJ	2024-03-15 13:30:00	2024-03-15 13:45:00				
451	Anguilla	Lloyd International Airport	NEW232	Scheduled	H25B	Boeing 737-8AL	PR-MYZ				United States	ningdale Republic Airport	KFRG	2024-03-15 12:30:00	2024-03-15 16:09:14			2024-03-15 12:40:00	
452	Anguilla	Lloyd International Airport	AA3419	Scheduled	E75L	Airbus A320-232	PR-MHK	United States	Envoy	ENY	United States	/iami International Airport	KMIA	2024-03-15 14:05:00	2024-03-15 17:05:00			2024-03-15 14:05:00	
453	Anguilla	Lloyd International Airport	Q3611	Scheduled	BNI	Boeing 787-9 Dreamliner					Netherlands	iliana International Airport	TNCM	2024-03-15 17:00:00	2024-03-15 17:10:00				
454	Anguilla	Lloyd International Airport	W12057	Scheduled	BNT	Boeing 737-89L	PR-YSG				Netherlands	iliana International Airport	TNCM	2024-03-15 17:00:00	2024-03-15 17:10:00				
455	Anguilla	Lloyd International Airport	Q3601	Scheduled	BNI	Boeing 737-89L					Netherlands	iliana International Airport	TNCM	2024-03-15 17:00:00	2024-03-15 17:10:00				
456	Anguilla	Lloyd International Airport		Scheduled	C68A	Airbus A320-232	PR-YYH	United States	NetJets	EJA	United States	ington New Castle Airport	KILG	2024-03-15 13:30:00	2024-03-15 17:18:00			2024-03-15 13:36:00	
457	Anguilla	Lloyd International Airport	AA3420	Scheduled	E75L		PR-YSA	United States	Envoy	ENY	United States	/iami International Airport	KMIA	2024-03-15 16:05:00	2024-03-15 18:50:00			2024-03-15 16:05:00	
458	Anguilla	Lloyd International Airport	Q3621	Scheduled	BNI	Boeing 737-85C					Netherlands	iliana International Airport	TNCM	2024-03-15 19:00:00	2024-03-15 19:10:00				
459	Anguilla	Lloyd International Airport	W12059	Scheduled	BNT	Boeing 737-89L					Netherlands	iliana International Airport	TNCM	2024-03-15 19:00:00	2024-03-15 19:10:00				
460	Anguilla	Lloyd International Airport	GJE1021	Scheduled		Boeing 737-89L					Puerto Rico	/arin International Airport	TJSJ	2024-03-15 19:00:00	2024-03-15 19:28:00			2024-03-15 19:00:00	
461	Anguilla	Lloyd International Airport	Q3641	Scheduled	BNI	Boeing 737-86N	UK72604				Netherlands	iliana International Airport	TNCM	2024-03-15 20:30:00	2024-03-15 20:40:00				

FIGURE 1.2 – Exemple de Dataframe obtenu (Arrivals)

3. **Jeux de données existants**

   Comme on peut le voir sur la figure I.2, il existe de nombreuses valeurs NaN. Pour certaines analyses, cela nous pose problème. Comme solution, nous avons utilisé des jeux de données existants qui ne concerne que les vols effectuées en 2015 aux Etats-Unis, cela nous permet de nous concentrer un peu plus sépcifiquement sur les données d'un seul pays, de ses aéroports et de ses compagnies aériennes.

1.3 Pré-traitement des données

Une fois les datasets crées et obtenus, nous avons du les prétaiter.

1. **Formattage des données**

   Consiste à reformatter le type de certaines colonnes. Par exemple pour la colonne **"Date"**, la mettre sous le format YYYY-MM-DD, et de type **datetime**. Pour les colonnes heures d'arrivée et de départ également ; dans le fichier .json obtenu grâce l'API [**api**], elles étaient sous format **timestamp**, il a donc fallu les mettre sous un nouveau format aussi de type **datetime**, on obtient **YYYY-MM-DD HH :MM :SS**

2. **Conversion des données**

   Pour le jeu de données téléchargé, nous avons par exemple convertit les données de la colonne **DISTANCE** des miles aux kilomètres.

3. **Compléments**

   En plus des colonnes existantes dans les dataframes des arrivées, départs et des compagnies aériennes , nous les avons complétés en ajoutant des colonnes qui définissent :

   — le type du vol (International/National)

   — le type de la compagnie aérienne (Low Cost/High Cost).

4. **Traduction**

   Dans le jeu de données des avis clients, on pouvait retrouver plusieurs langues différentes, nous avons donc utilisé la librairie **GoogleTranslator** afin de traduire tous les commentaires en anglais.

Chapitre 2

Analyse des données

Cette partie détaille le protocole d'exploration des données mis en place. Les librairies utilisées sont :

— Pour la manipulation et analyse des données : **Pandas**

— Pour la visualisation **Seaborn** et **Matplotlib**

2.1 Analyse des compagnies aériennes

2.1.1 Chargement et séparations des données

Comme mentionné précédemment, les données ont été chargées et nettoyées en supprimant les colonnes avec plus de 60% et 65% de valeurs manquantes (voir tableau 2.1) pour les jeux de données des arrivées et des départs respectivement. De plus, nous avons séparé les données entre les compagnies low-cost et full-service afin de comparer ces deux types de modèles économiques.

Colonne	**Pourcentage de NaN**
Flight Duration	98.96%
Real Arrival	98.95%
Estimated Arrival	90.68%
Real Departure	89.43%

Table 2.1 – Nombre et pourcentage de valeurs manquantes par colonne supprimée

Suite à cela, nous avons créé un jeu de données regroupant les arrivées et les départs des compagnies low-cost et full-service pour faciliter l'analyse. Les colonnes finales sont :

Colonnes finales	Colonnes finales (suite)
Country	Scheduled Departure
Airport	Scheduled Arrival
Flight identification	Estimated Departure
Flight Status	Destination Country
Aircraft Code	Destination Airport Name
Aircraft Name	Destination Airport Code
Aircraft Registration	Departure Airport Code
Aircraft owner (Country)	Departure Airport Name
Airline Name	Airline Code

TABLE 2.2 – Colonnes finales du jeu de données

2.1.2 Comparaison des statistiques de vol par type de service

Au total, nous constatons qu'il y a nettement plus de vol de type full-service que de type low-cost (4,4 fois plus). Cependant, cela est normal étant donné qu'il y a moins de compagnies low-cost. C'est pourquoi nous allons plutôt nous baser sur la fréquence de vol par type de compagnie. Dans ce cas, il est clair que les compagnies low-cost effectuent beaucoup plus de vol, avec une médiane de 10 vol contre 3 pour les compagnies full-service.

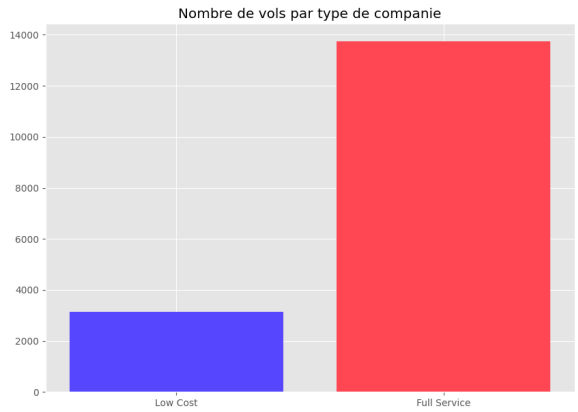


FIGURE 2.1 – Nombre de vol par type de service

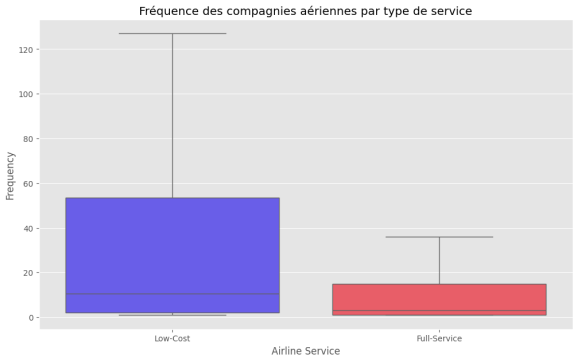


FIGURE 2.2 – Fréquence de vol par type de service (sans outliers)

Parmi toutes les compagnies, le top 5 des compagnies avec le plus de fréquence de vol est representé dans le tableau 2.3, et le top 30 dans le graphique 2.3

Nom de la compagnie	Fréquence	Type de service
Southwest Airlines	829	Low-Cost
IndiGo	725	Low-Cost
Delta Air Lines	675	Full-Service
SkyWest Airlines	675	Full-Service
American Airlines	665	Full-Service

TABLE 2.3 – Top 5 des compagnies avec le plus de fréquence de vol

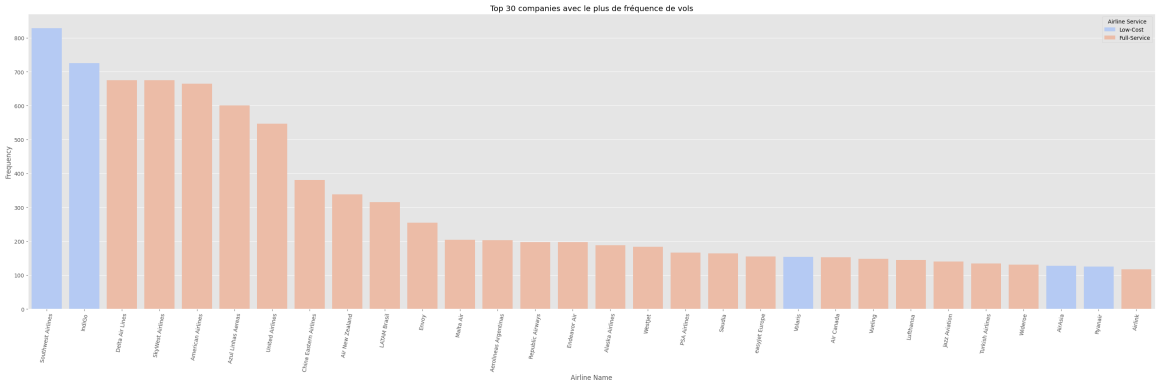


FIGURE 2.3 – Top 30 companies avec le plus de fréquence de vol

2.1.3 Comparaison des types d'avions utilisé par type de services

Nous regroupons les avions en deux catégories principales : les avions long-courrier et moyen-courrier.

Notre analyse montre clairement que les types d'avions les plus fréquemment utilisés sont les avions moyen-courrier (Boeing 737-800, Airbus A320) et cela pour les deux types de compagnies.

Cependant, il est à noter que les compagnies low-cost utilisent très rarement des avions long-courrier. Nous n'en recensons que 7 sur 84 qui utilisent des Airbus A330, des Boeing 777 et des Boeing 787 : Air Europa, Jetstar, Thai AirAsia X, Scoot, AirAsia X, Lion Air, et Jin Air, contre 90 pour les compagnies full-service.

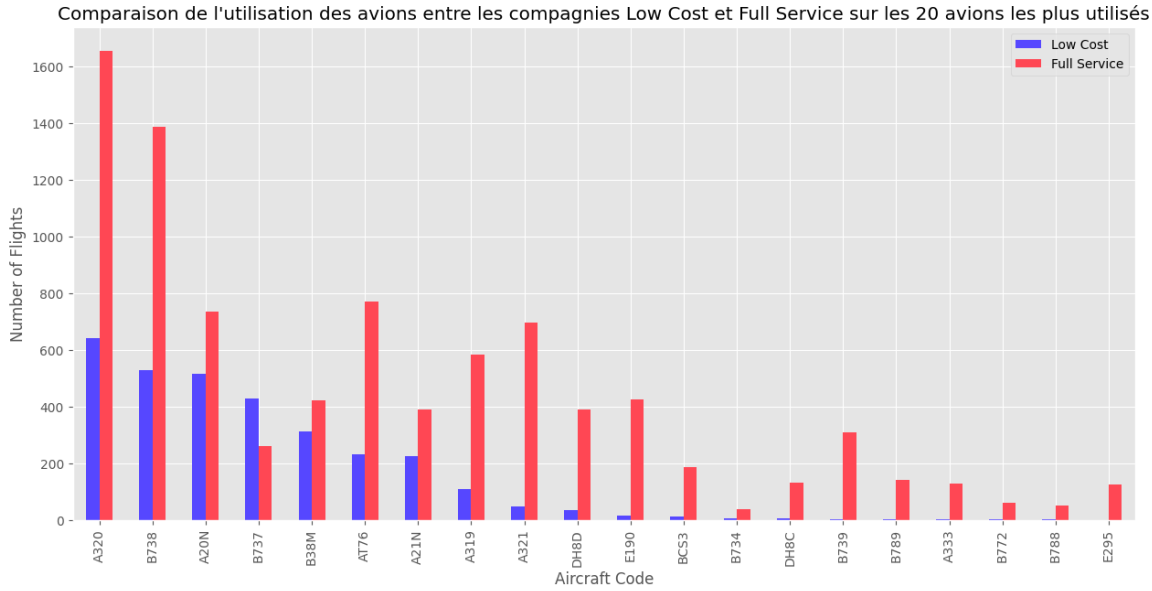


FIGURE 2.4 – Comparaison de l'utilisation des avions entre les compagnies Low Cost et Full Service sur les 20 avions les plus utilisés

2.1.4 Comparaison des aéroports et destinations desservis par les deux types de compagnies

Analyse des destinations desservis

Étant donné que les compagnies low-cost utilisent moins d'avions long-courrier, il est prévisible qu'elles desservent moins de pays, ce que confirme la carte du monde ci-dessous.

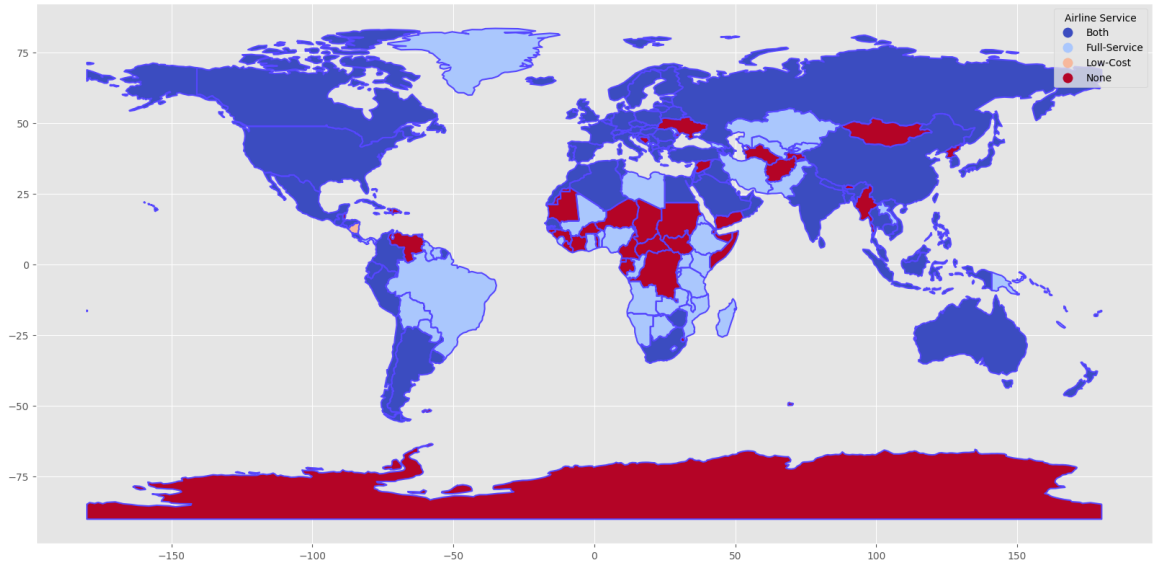


FIGURE 2.5 – Pays desservis par les types de compagnies

Remarque : Certaines zones apparaissent en rouge, mais elles sont bien desservies par des compagnies full-service. Cela est dû au fait que les données ont été collectées sur une courte période.

Analyse des aéroports desservis

Les compagnies full-service desservent bien plus d'aéroports internationaux que les compagnies low-cost, ce qui est illustré par le tableau ci-dessous.

Type de compagnie	Nombre d'aéroports internationaux desservis
Low-Cost	239
Full-Service	438

TABLE 2.4 – Nombre d'aéroports internationaux desservis par type de compagnie

2.1.5 Comparaison des retards et des distances parcourues aux États-Unis

Pour faire cette comparaison, nous nous sommes basés sur le marché des États-Unis, étant donné que c'est notre marché cible et qu'il y avait un manque d'informations dans les jeux de données initiaux.

Tout d'abord, nous constatons que la médiane des distances parcourues par les compagnies low-cost est plus élevée (1121 kilomètres contre 1018 kilomètres). Ainsi, malgré le manque de vols long-courrier, le fait que les compagnies low-cost volent bien plus fréquemment permet de compenser. En effet, la distance parcourue est plus importante, ce qui semble indiquer un compromis entre vols long-courrier et fréquence de vol, permettant des économies pour les compagnies low-cost.

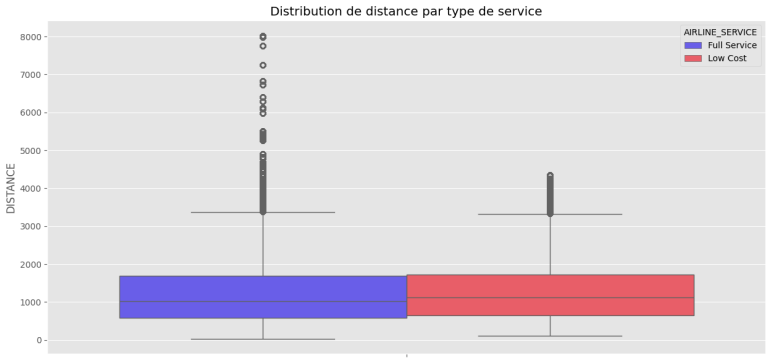


FIGURE 2.6 – Distribution des distances parcourues par type de service

Ensuite, nous observons que les compagnies low-cost (sans prendre en compte les valeurs aberrantes pour la visualisation), ont tendance à avoir plus de retards. En effet, la médiane se trouve à 0 (arrivée à l'heure) contre -2 minutes (arrivée en avance) pour les compagnies full-service. Cela peut sembler être un petit décalage non significatif au premier abord, mais nous allons effectuer un test de significativité pour confirmer ce résultat qui prendra en compte les valeurs abérrantes cette fois ci.

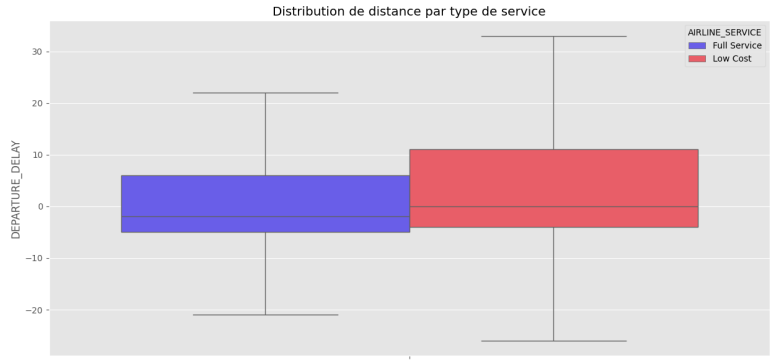


FIGURE 2.7 – Distribution des retards de départ par type de service (sans valeurs aberrantes)

Tests statistiques

Pour évaluer la significativité des différences observées, nous avons réalisé les tests statistiques suivants :

- **Test de Bartlett** pour l'homogénéité des variances.
  - Résultat : le test donne une p-valeur de 0.0, indiquant que la variable "DEPARTURE_DELAY" ne possède pas une variance homogène.
- **Test de Shapiro** pour vérifier la normalité des retards.
  - Résultat : le test donne une p-valeur de 1.0, confirmant que la variable "DEPARTURE_DELAY" suit une loi normale.
- **ANOVA de Welch** pour comparer les groupes, et le **Test de Games-Howell** pour les comparaisons post-hoc.
  - Approche : étant donné l'absence d'homogénéité des variances, nous avons utilisé l'ANOVA de Welch, plus robuste que l'ANOVA classique.
  - Résultat :

Source	**ddl1**	**ddl2**	**F**	**p-unc**	**np2**
SERVICE_AÉRIEN	1	3520523.364	4036.439	0.000	0.001

TABLE 2.5 – Résultats de l'ANOVA de Welch pour la comparaison des retards de départ par type de service

TABLE 2.6 – Comparaison des retards de départ par type de service

		Moyenne					
Groupe A	Groupe B	A	B	Diff	SE	T	ddl
Service complet	Compagnie low-cost	8.778	10.822	-2.045	0.032	-63.533	3520523.364

Les résultats indiquent une p-value de 0, démontrant ainsi une différence significative entre les compagnies low-cost et full-service en termes de retards.

2.2 Analyse des avis clients

Après avoir nettoyé les données et retiré les colonnes avec un pourcentage élevé de valeurs manquantes (voir tableau 2.7), nous avons remplacé les valeurs manquantes des autres colonnes par les notes de la colonne "General Stars". Cette approche est apparue comme la plus judicieuse. Ainsi, nous avons pu analyser les avis des clients sur les aéroports. Voici les principaux points d'intérêt issus de notre analyse

Column	Pourcentage de NaN
Immigration/customs	74.87%
Baggage claim	73.01%
Security check	52.15%
Lounge	49.80%

TABLE 2.7 – Pourcentage de valeurs manquantes par colonne supprimée

2.2.1 Detection d'anomalie

En analysant les avis, nous avons détecté des anomalies où les utilisateurs ont attribué des notes générales incohérentes par rapport aux sous-notes. Par exemple, des avis avec une note générale de 5 étoiles mais des sous-notes de 1 étoile. Nous avons corrigé ces incohérences pour garantir une analyse plus précise.

	Country	Airport	Date	Content	General Stars	Getting to the Airport	Terminal facilities	WiFi	Food and retail services	Lounge	Immigration/customs	Baggage claim	Security check
28	Algeria	Algiers Houari Boumediene Airport	2015-01-15	One of the nicest in Africa, very bright and c...	4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	NaN
36	Algeria	Tlemcen Zenata Airport	2014-11-21	L'aéroport de Tlemcen est un petit aéroport re...	5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	NaN
46	Angola	Luanda Quatro de Fevereiro Airport	2014-08-22	You are lucky if you are not arrested by the a...	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	NaN

FIGURE 2.8 – Exemple d'avis client incohérent

2.2.2 Nombre d'avis positif et negatif

Ensuite, nous avons voulu évaluer l'équilibre entre les avis positifs et négatifs. Les notes allant de 0 à 5, nous avons considéré les avis négatifs comme ceux ayant une note inférieure à 3, et les avis positifs comme ceux ayant une note supérieure ou égale à 3.

Nous avons observé la distribution globale des notes pour visualiser cet équilibre. Comme le montre la figure 2.9, les notes ne suivent pas de loi normale, et la figure 2.10 montre qu'il y a clairement beaucoup plus d'avis positifs que négatifs en général. Cependant, cette tendance peut différer selon les pays, comme illustré par la figure 2.11.

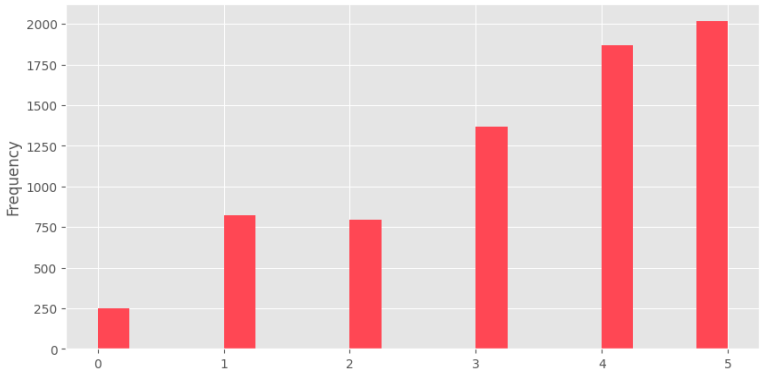


FIGURE 2.9 – Distribution des notes

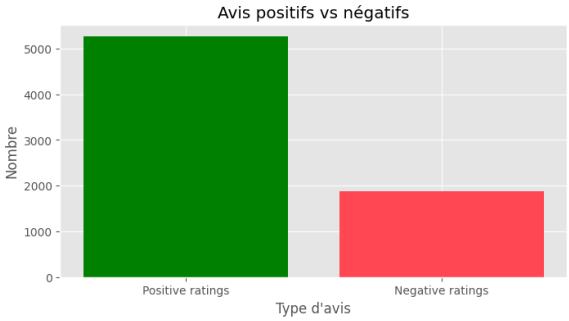


FIGURE 2.10 – Nombre de types d'avis

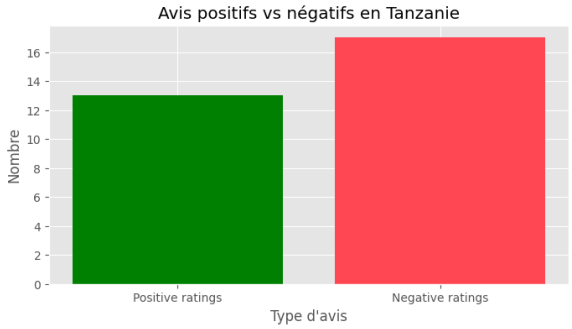


FIGURE 2.11 – Nombre de types d'avis par pays

Pour finir, nous avons recensé les aéroports les mieux notés et les pires notés de notre jeu de données.

Top 5 des aéroports les moins bien notés	**Note moyenne**
Cuneo Levaldigi Airport (Italie)	1.00
Mount Pleasant Airport (Îles Falkland)	0.80
Metz-Nancy-Lorraine Airport (France)	0.75
Batumi International Airport (Géorgie)	0.67
Reggio Calabria Airport (Italie)	0.67

TABLE 2.8 – Top 5 des aéroports les moins bien notés

Top 5 des aéroports les mieux notés	**Note moyenne**
Rostov-on-Don Platov International Airport (Russie)	5.00
Huatulco International Airport (Mexique)	5.00
Monteria Los Garzones Airport (Colombie)	5.00
Burnie Airport (Australie)	5.00
Warsaw Radom Airport (Pologne)	5.00

TABLE 2.9 – Top 5 des aéroports les mieux notés

2.2.3 Analyse de corrélation

Nous avons voulu analyser la corrélation entre les notes générales et les sous-notes. Comme le montre la figure 2.12, certaines sous-notes sont fortement corrélées avec la note générale, par exemple les notes de "Terminal facilities".

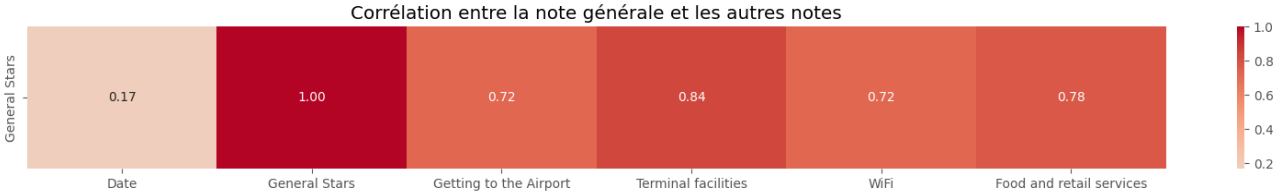


FIGURE 2.12 – Corrélation entre la note générale et les autres notes

2.2.4 Analyse des topics des avis

Nous avons ensuite voulu analyser quels étaient les facteurs impactant le plus les notes en plus de l'analyse de corrélation afin de combler le manque de données. Nous avons effectué une analyse des sujets(topics) des commentaires en utilisant la méthode LDA (Latent Dirichlet Allocation).

Grâce à cette analyse, nous avons identifié et interprété quatre principaux sujets abordés dans les avis positifs et négatifs :

— **Avis positifs** : Sécurité, Facilité d'accès, Terminale moderne, et Services de nourriture.

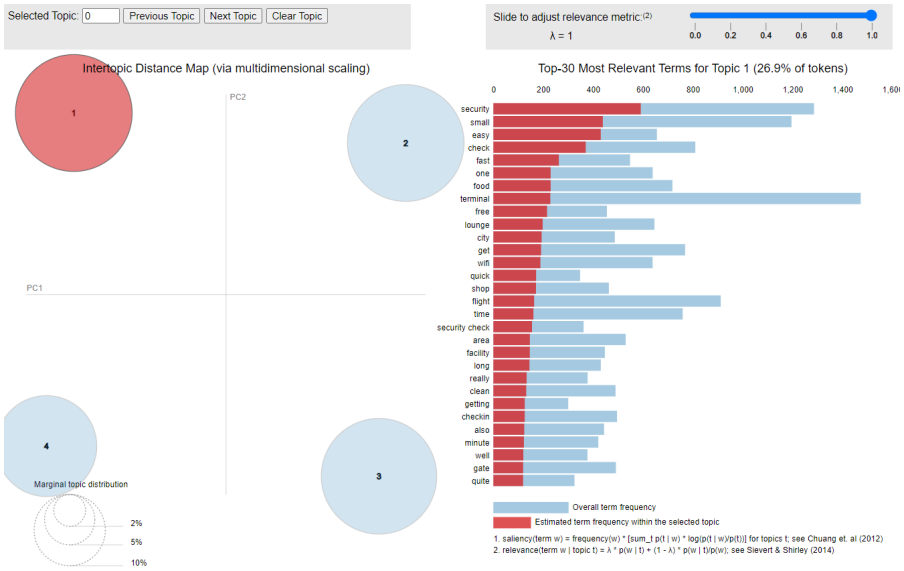


FIGURE 2.13 – Visualisation de la LDA sur le 1er topic des avis positifs

— **Avis négatifs** : Temps d'attente, Sécurité, Mauvais Staff, Manque de propreté.

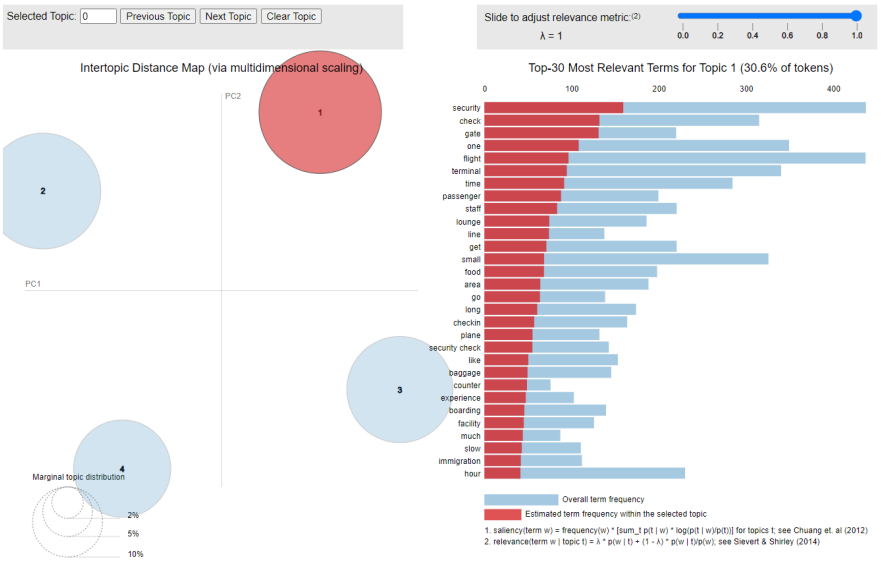


FIGURE 2.14 – Visualisation de la LDA sur le 1er topic des avis negatifs

2.2.5 Zoom sur les États-Unis

Étant donné que notre marché cible sont les États-Unis, nous avons effectué les mêmes analyses spécifiques à ce pays. Les résultats sont les suivants :

Distribution des notes :

On compte 748 avis positifs pour 133 avis negatifs, la distribution des notes est representé sur la figure ci-dessous :

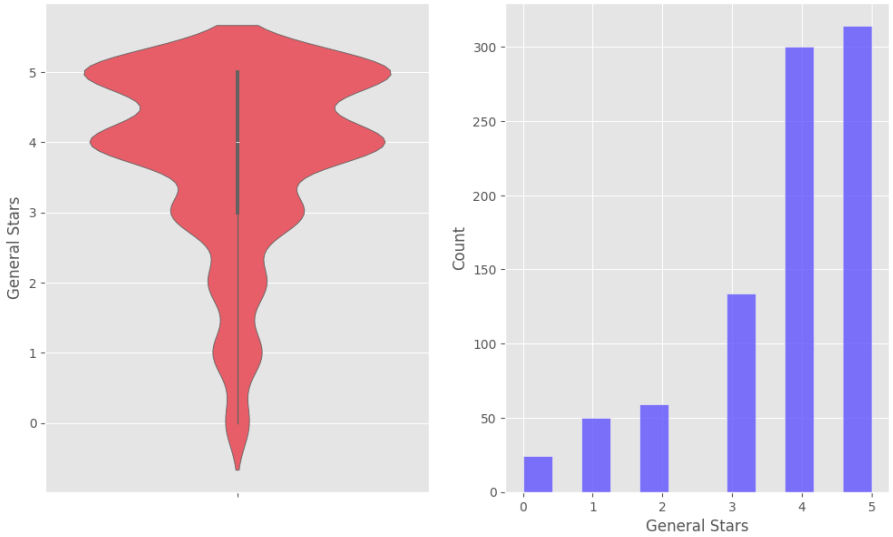


FIGURE 2.15 – Distribution des notes au États Unis

Analyse sur les aéroports :

Les résultats montrent les aéroports les mieux notés et les moins bien notés :

Top 3 des aéroports les mieux notés aux USA	Note
San Diego International Airport	4.5
Portland International Airport	4.4
Honolulu International Airport	4.3

TABLE 2.10 – Top 3 des aéroports les mieux notés aux USA

Top 3 des aéroports les moins bien notés aux USA	Note
LaGuardia Airport	2.5
Newark Liberty International Airport	2.3
Cleveland Hopkins International Airport	2.2

TABLE 2.11 – Top 3 des aéroports les moins bien notés aux USA

Analyse des topics des avis :

— **Avis positifs** : Sécurité, Facilité d'accès, Terminale moderne, Personnel amical

— **Avis négatifs** :Temps d'attente, Sécurité, File d'attente, Propreté

2.3 Analyse des retards

Cette analyse a été réalisé sur le dataset contenant uniquement les données de vol des Etats-Unis sur une année entière (2015). Afin de réduire la taille du dataset, nous nous concentrons que sur les statistiques du mois d'Avril. Après le pré-traitement effectué (Partie I.3), nous avons d'abord commencer par analyser le pourcentage de valeurs manquantes dans le jeu de données en calculant le pourcentage de remplissage du dataset. Sept sur treize colonnes était remplies à 100% et les autres colonnes restantes à 99-98 %. Malgré qu'il y ait quelques valeurs manquantes mais en majorité le jeu de données est bien rempli. Nous avons donc supprimé le peu de valeurs NaN existantes.

Nous avons ensuite pu entamer le travail d'analyse :

2.3.1 Analyse statistique descriptive

1. **De tous le jeu de données**
   Grâce à la fonction **describe** en Pandas, nous avons pu obtenir les valeurs moyennes, minimum, maximum de chaque colonne. Nous remarquons que pour les colonnes **DEPARTURE_DELAY** et **ARRIVAL_DELAY** qui décrivent le retard en minutes que le vol à fait respectivement au décollage et à l'arrivée ont un retard moyen respectifs de 7,72 min et de 3,16 min dans tous le dataset. De plus on note que 75% des vol font un retard d'environ 6 et 7 min au décollage et à l'arrivée.

Les valeurs aberrantes que l'on peut voir également sont des retards au décollage de 24,9 heures et à l'arrivée de 25,9 heures.

2. **En fonction des compagnies aériennes**

   Dans le jeu de données, nous retrouvons 14 compagnies aériennes différentes.

   Voici les résultats obtenu après observation des statistiques (les nombres négatifs démontrent un temps d'avance) :

	Plus de retard	**Moins de retard**	**Pire retard**	**Meilleure avance**
Durée (min)	12	-2	25.7	68
Compagnie aérienne	United Air Lines Inc.	Hawaiian Airlines Inc.	American Airlines Inc.	American Airlines Inc.

TABLE 2.12 – Analyse statistique des retards en fonction des compagnies aériennes

De plus en calculant le pourcentage de retards moyen par compagnies et le pourcentage de vol que chaque compagnies effectuent, on obtient les camemberts suivants :

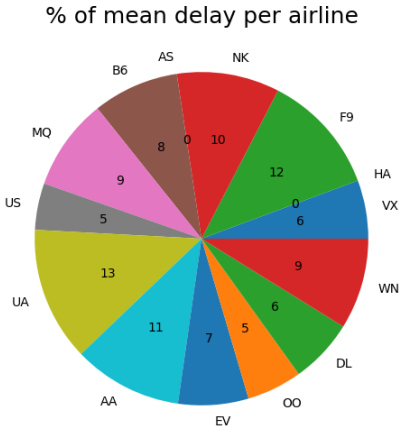


FIGURE 2.16 – Retards moyens (min) par compagnies

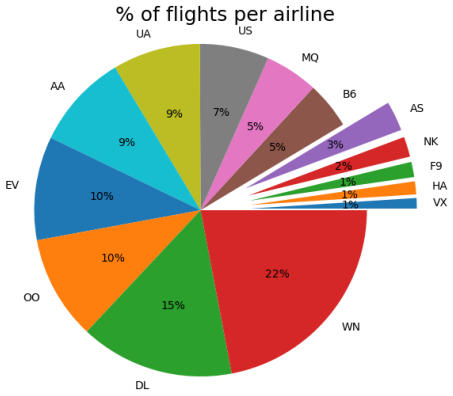


FIGURE 2.17 – % de vol effectués par compagnies

On remarque que la compagnie aérienne WN (Southwest Airlines Co.) est la compagnie aérienne qui effectue le plus de vol pourtant ce n'est pas celle qui commet la plus grande moyenne de retards (elle est classée 5ème parmi toutes les compagnies).

Toutefois le premier camembert montre qu'il y a très peu de différences entre les compagnies aériennes pour les retards commis, à l'exception des compagnies Hawaiian Airlines (HA) et Alaska Airlines (AS) qui ont une moyenne proche de 0, le reste des compagnies ont une moyenne qui reste relativement faible entre 7 et 13 min.

On remarque également sur le graphe ci-dessous (figure 2.3) que pour la majorité des compagnies aériennes, le retard qu'elles ont effectué au décollage est rattrapée lors du vol. Elles ajustent la durée du vol afin qu'il n'y ait pas de conséquences sur l'heure d'arrivée,

sauf pour les compagnies AS : Alaska Airlines Inc. et HA : Hawaiian Airlines Inc. qui respectivement arrivent en avance et décollent avant l'heure prévue.

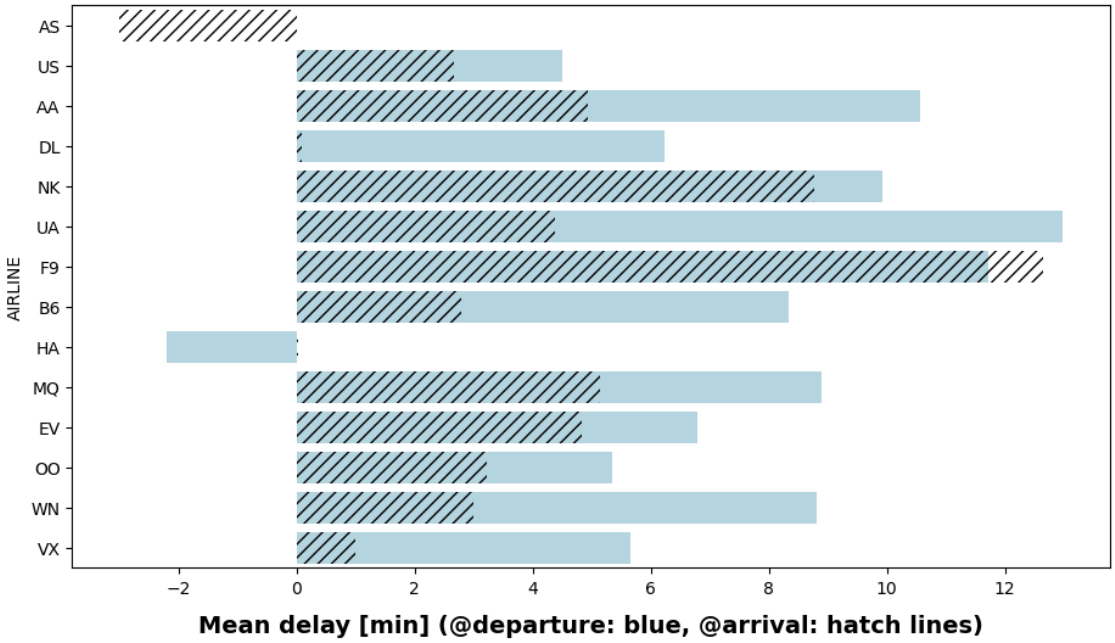


FIGURE 2.18 – Délais au décollage vs à l'arrivée

3. **En fonction des aéroports**

Nous avons dans un premier temps afficher les statistiques de tous les aéroports. Le retard moyen au décollage pour le mois d'avril en tous les aéroports est d'environ cinq minutes ce qui est acceptable. Toutefois, l'aéroport Jack Brooks Regional (BPT) qui connaît le plus de délais avec environ 31,78 minutes d'attente. Mais cette analyse seule ne suffit pas à classer l'aéroport BPT comme étant le pire. Afin d'approfondir cette analyse, on étudie également les statistiques des aéroports en fonction de chaque compagnie aérienne. A première vue, on pourrait classe l'aéroport Gunnison–Crested Butte Regional (GUC) comme étant le meilleur, avec des vols qui décollent en avance d'environ 14 minutes, et l'aéroport Rafael Hernández (BQN) comme étant le pire, avec des retards allant jusqu'à 74 minutes. Mais le nombre de vols à destination et au départ de ces aéroports n'est pas le même, le résultat peut être biaisé, on normalise donc les données et on obtient ce classement :

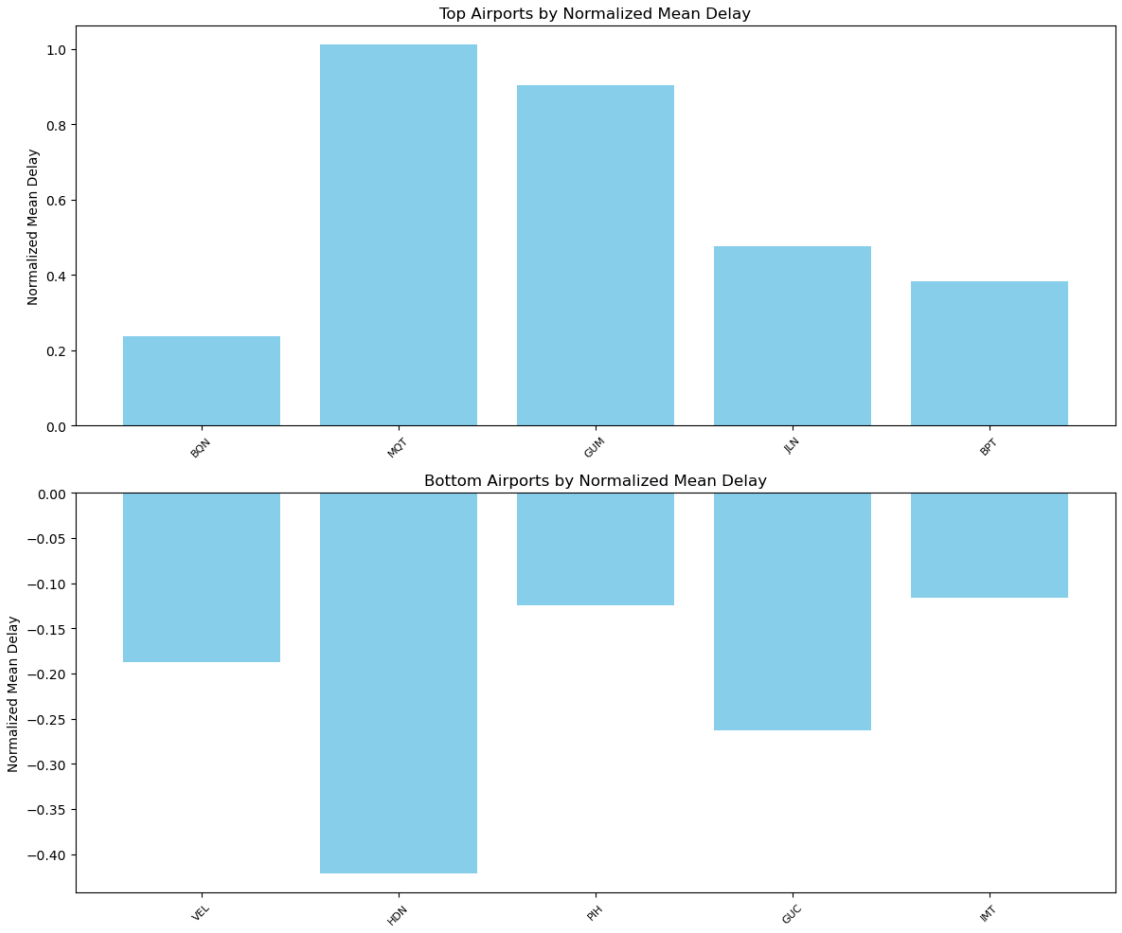


FIGURE 2.19 – Classement des aéroports en fonctions des délais au décollage

Ce classement confirme la position de l'aéroport GUC comme étant parmi les meilleurs (dans le top 3) et celle de l'aéroport BQN comme l'un des pires (après les aéroports Guam International (GUM) et Sawyer International (MQT).

On fait exactement les mêmes analyses sur les délars à l'arrivée, et les résultats sont similaires (l'aéroport Norfolk International, ORF est le pire dans ce classement).

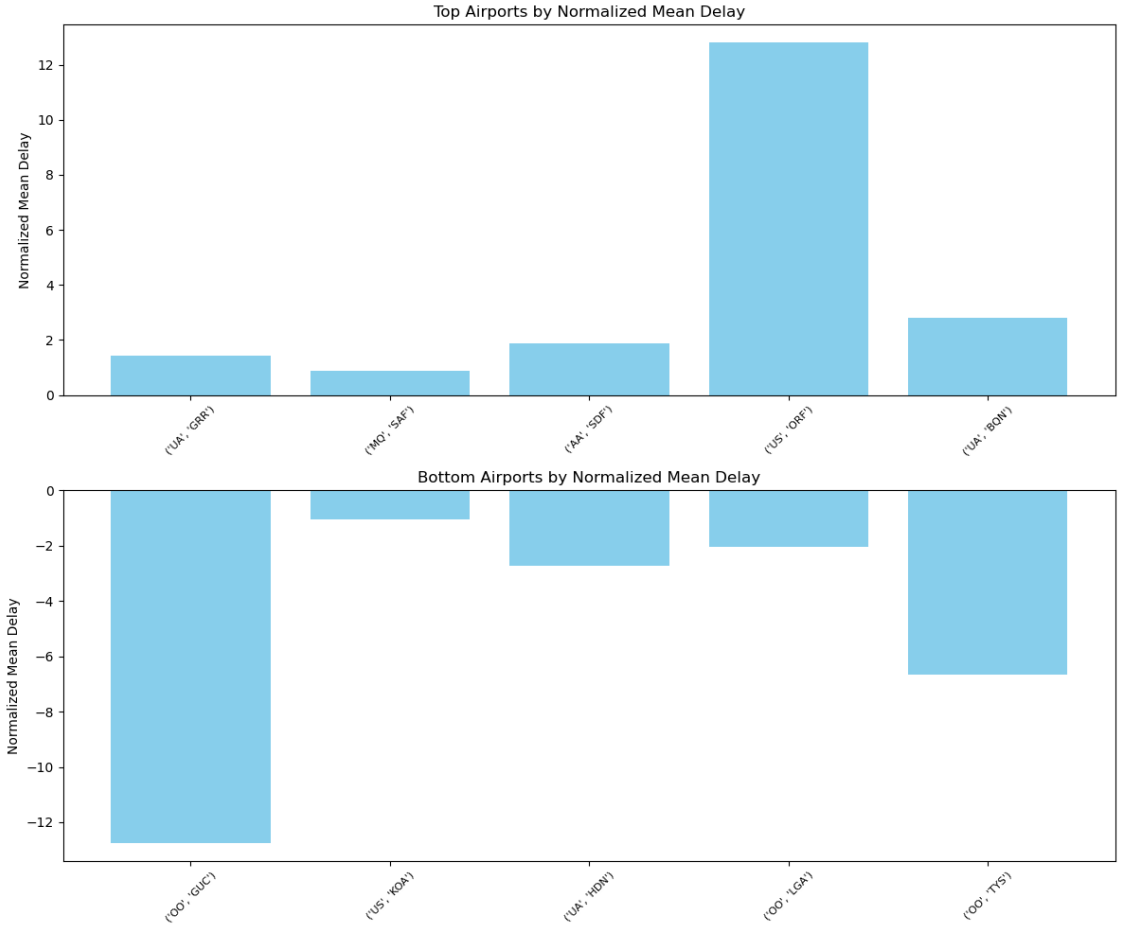


FIGURE 2.20 – Classement des aéroports en fonctions des délais à l'arrivée

2.3.2 Visualisation des distributions

Nous avons réalisé des diagrammes en boîtes et en violons afin de visualiser la dispersion des délais de décollage pour chaque compagnie aérienne à l'aide des fonctions **violonplot** et **boxplot**. Ce qui en ressort c'est que la majorité des compagnies ont un retard approximativement acceptable (comme ce que l'on a vu dans la partie 2.3.1). De plus les diagrammes nous ont permis de détecter certaines valeurs aberrantes par exemple des retards de plusieurs heures qui sont bien évidemment de rares occurences.

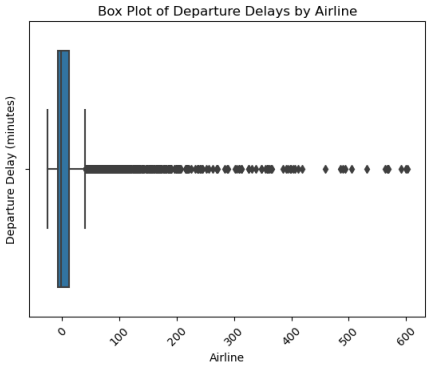


FIGURE 2.21 – Diagramme en boîtes pour l'aéroport F9

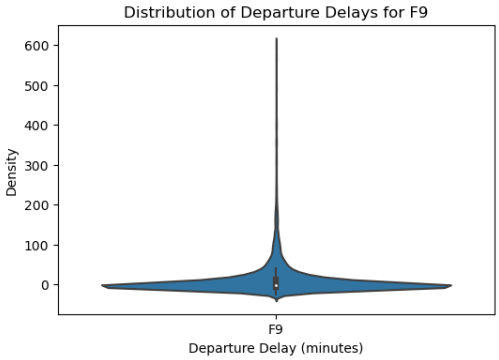


FIGURE 2.22 – Diagramme en violon pour l'aéroport F9

2.3.3 Analyse des tendances temporelles

Pour cette partie de l'analyse, on reprends le jeu de données entier, car on veut étudier la variation des retards selon les jours, de la semaine, les mois de l'année et selon les saisons.Afin de visualiser les résultats, on utilise des diagrammes à barres. On visualise deux choses :

— retards moyens

— proportion de vols

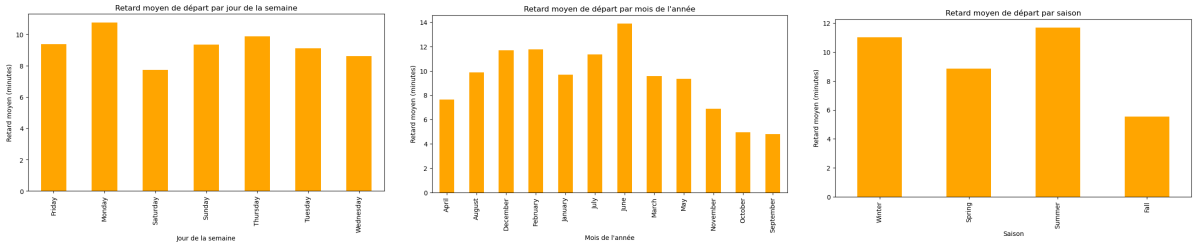


FIGURE 2.23 – Retards moyens par jours, mois, saisons

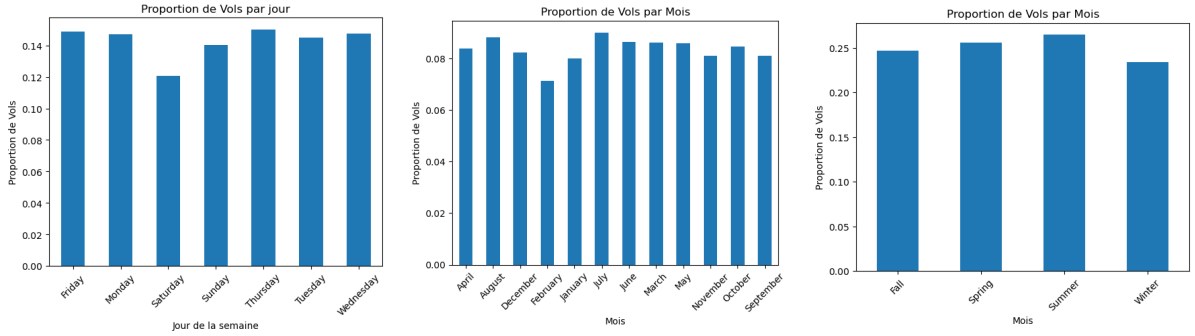


FIGURE 2.24 – Proportion de vols par jours, mois, saisons

La conclusion qu'on tire cette analyse est que pour les jours de la semaine où il y a le plus de retards, ce sont les jours où il y a le plus de vols (soit le lundi). De plus c'est durant l'été que le plus de retards sont enregistrés (surtout le mois de juin dans l'analyse par mois) ce qui

pourrait être expliqué par le nombre de vols à cette période qui est plus élevée par rapport au nombre tout au long de l'année.

La deuxième saison qui enregistre le plus de retards est l'hiver (surtout le mois de février et décembre dans l'analyse par mois), toutefois ce n'est pas la saison où le plus de vols à été enregistré après l'été mais cela peut s'expliquer par les conditions météorologiques.

2.3.4 Corrélation entre les variables

On étudie ici, les liens de colinéarité entre les variables, grâce à

1. **Matrice de corrélation**

   D'après la matrice de corrélation obtenues, on peut en tirer les informations suivantes :

   — Corrélations fortes :

   Les variables **DEPARTURE_DELAY** et **ARRIVAL_DELAY** présentent une forte corrélation (0.87) ce qui est attendu puisque les retards au décollage entraîne souvent des retards à l'arrivée. De même pour les variables **SCHEDULED_TIME** et **ELAPSED_TIME** indiquant que le temps de vol prévu est souvent proche du temps de vol réel.

   — Corrélation entre les mois et les jours de la semaine

   Les mois et les jours de la semaine montrent généralement une faible corrélation avec **DEPARTURE_DELAY**, à l'exception de certains mois spécifiques comme les mois de Juin, Décembre et Février qui montrent une légère corrélation positive (ce qui confirme ce qui as été dit en partie 2.3.3). Pour les jours de la semaine, il y a très peu de corrélations avec la variable **DEPARTURE_DELAY**, avec des valeurs allant de -0.01 à 0.03 ce qui implique que le jour de la semaine n'influence pas significativement les retards au décollage (sauf pour le lundi).

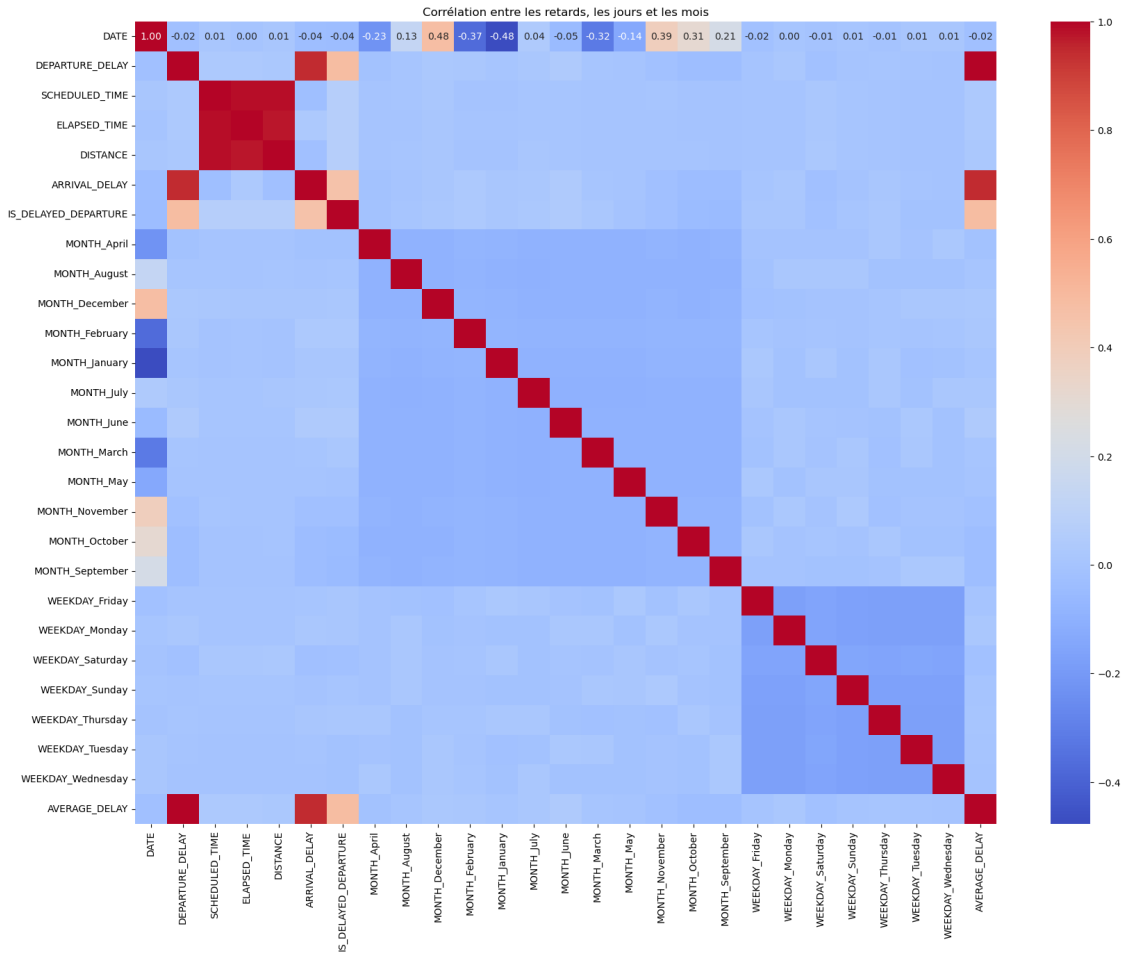


FIGURE 2.25 – Matrice de corrélation

2. **Régression multiple**

   La régression multiple est également une technique statistique qui permet d'analyser la relation entre une variable dépendantes et plusieurs variables indépendantes. Les résultats avec la variable **DEPARTURE_DELAY** montrent que le coefficient constant est de 6,5669 signifiant qu'en l'absence de toutes les autres variables, le retard moyen est d'environ 6,57 minutes. On confirme également que les mois de décembre, février et juin ont des retards plus élevées (coefficients positifs significatifs) contraitement aux mois de mai, septembre ou encore octobre. Pour les jours de la semaine également, le lundi a un petit effet positif. Enfin, le coefficient relatif à la distance de vols est de 0.0009 et est statistiquement significatif. Pour la même analyse au niveau des arrivées, on obtient les mêmes résultats sauf pour la distance qui as un coefficient pratiquement nul et négatif (-0.0012) et est statistiquement non significatif, indiquant que la distance parcourue n'influence pas significativement les retards à l'arrivée, car comme on l'a vu dans la figure 2.18, la majorité des vols qui décollent en retard arrivent à le rattraper légèrement à l'arrivée quelque soit la distance.

OLS Regression Results

Dep. Variable:	DEPARTURE_DELAY	R-squared:	0.006
Model:	OLS	Adj. R-squared:	0.006
Method:	Least Squares	F-statistic:	2011.
Date:	Mon, 20 May 2024	Prob (F-statistic):	0.00
Time:	04:06:34	Log-Likelihood:	-2.8706e+07
No. Observations:	5714008	AIC:	5.741e+07
Df Residuals:	5713989	BIC:	5.741e+07
Df Model:	18		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.5669	0.068	96.435	0.000	6.433	6.700
MONTH_August	2.2145	0.074	29.822	0.000	2.069	2.360
MONTH_December	4.0545	0.076	53.689	0.000	3.906	4.202
MONTH_February	4.1649	0.078	53.148	0.000	4.011	4.319
MONTH_January	2.0992	0.076	27.595	0.000	1.950	2.248
MONTH_July	3.6825	0.074	49.872	0.000	3.538	3.827
MONTH_June	6.1852	0.075	82.853	0.000	6.039	6.331
MONTH_March	1.9164	0.075	25.658	0.000	1.770	2.063
MONTH_May	1.7584	0.075	23.515	0.000	1.612	1.905
MONTH_November	-0.7899	0.076	-10.411	0.000	-0.939	-0.641
MONTH_October	-2.6820	0.075	-35.754	0.000	-2.829	-2.535
MONTH_September	-2.8133	0.076	-37.095	0.000	-2.962	-2.665
WEEKDAY_Monday	1.3290	0.057	23.481	0.000	1.218	1.440
WEEKDAY_Saturday	-1.7218	0.060	-28.888	0.000	-1.839	-1.605
WEEKDAY_Sunday	-0.0545	0.057	-0.952	0.341	-0.167	0.058
WEEKDAY_Thursday	0.4919	0.056	8.740	0.000	0.382	0.602
WEEKDAY_Tuesday	-0.3353	0.057	-5.897	0.000	-0.447	-0.224
WEEKDAY_Wednesday	-0.7607	0.057	-13.448	0.000	-0.872	-0.650
DISTANCE	0.0009	1.57e-05	56.499	0.000	0.001	0.001

Omnibus:	7768268.264	Durbin-Watson:	1.857
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3746610706.424
Skew:	7.620	Prob(JB):	0.00
Kurtosis:	127.516	Cond. No.	2.06e+04

FIGURE 2.26 – Régression mutliple retard moyen au décollage

OLS Regression Results

Dep. Variable:	ARRIVAL_DELAY	R-squared:	0.008
Model:	OLS	Adj. R-squared:	0.008
Method:	Least Squares	F-statistic:	2520.
Date:	Mon, 20 May 2024	Prob (F-statistic):	0.00
Time:	04:10:43	Log-Likelihood:	-2.9058e+07
No. Observations:	5714008	AIC:	5.812e+07
Df Residuals:	5713989	BIC:	5.812e+07
Df Model:	18		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.8322	0.072	66.711	0.000	4.690	4.974
MONTH_August	1.5354	0.079	19.437	0.000	1.381	1.690
MONTH_December	3.0066	0.080	37.429	0.000	2.849	3.164
MONTH_February	5.1496	0.083	61.777	0.000	4.986	5.313
MONTH_January	2.6699	0.081	32.994	0.000	2.511	2.828
MONTH_July	3.3056	0.079	42.086	0.000	3.152	3.460
MONTH_June	6.4697	0.079	81.473	0.000	6.314	6.625
MONTH_March	1.7696	0.079	22.273	0.000	1.614	1.925
MONTH_May	1.4155	0.080	17.797	0.000	1.260	1.571
MONTH_November	-2.0478	0.081	-25.375	0.000	-2.206	-1.890
MONTH_October	-3.9305	0.080	-49.260	0.000	-4.087	-3.774
MONTH_September	-3.9071	0.081	-48.433	0.000	-4.065	-3.749
WEEKDAY_Monday	1.2104	0.060	20.104	0.000	1.092	1.328
WEEKDAY_Saturday	-2.8586	0.063	-45.089	0.000	-2.983	-2.734
WEEKDAY_Sunday	-0.7278	0.061	-11.945	0.000	-0.847	-0.608
WEEKDAY_Thursday	0.9064	0.060	15.140	0.000	0.789	1.024
WEEKDAY_Tuesday	-0.5839	0.060	-9.657	0.000	-0.702	-0.465
WEEKDAY_Wednesday	-0.8882	0.060	-14.763	0.000	-1.006	-0.770
DISTANCE	-0.0010	1.67e-05	-61.104	0.000	-0.001	-0.001

Omnibus:	7117631.091	Durbin-Watson:	1.835
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2378742750.117
Skew:	6.540	Prob(JB):	0.00
Kurtosis:	102.096	Cond. No.	2.06e+04

FIGURE 2.27 – Régression mutliple retard moyen à l'arrivée

2.3.5 Principal Component Analysis (PCA)

En réduisant la dimensionalité à deux composantes principales, celles qui expliques le plus la variance du jeu de données, ça nous as permis de visualiser dans un cercle de corrélation la force et la direction des relations entre les variables originales et les axes principaux.

1. Composantes principales :

   La première composante principapel (**SCHEDULED_TIME** explique 59,26% de la variance du dataset, elle capture donc bien l'information principale des données.

   La deuxième composante PC2 (**ARRIVAL_DELAY** explique quand à elle 38,88% de la variance.

2. Corrélations avec les composantes :

   La variable **DEPARTURE_DELAY** est très proche de l'axe vertical ce qui indique une forte corrélation avec la variable **ARRIVAL_DELAY**, cela souligne une transmission ou une continuation des retards de l'un à l'autre.

   Pour les variables **ELAPSED_TIME** et **DISTANCE**, sont positionnées proches de l'axe horizontal et très près de l'origine, ce qui suggère une corrélation faible avec les deux composantes principales (PC1 et PC2). Cela peut indiquer que ces variables ont un impact moins direct ou moins variable sur les retards de départ et d'arrivée par rapport à d'autres facteurs.

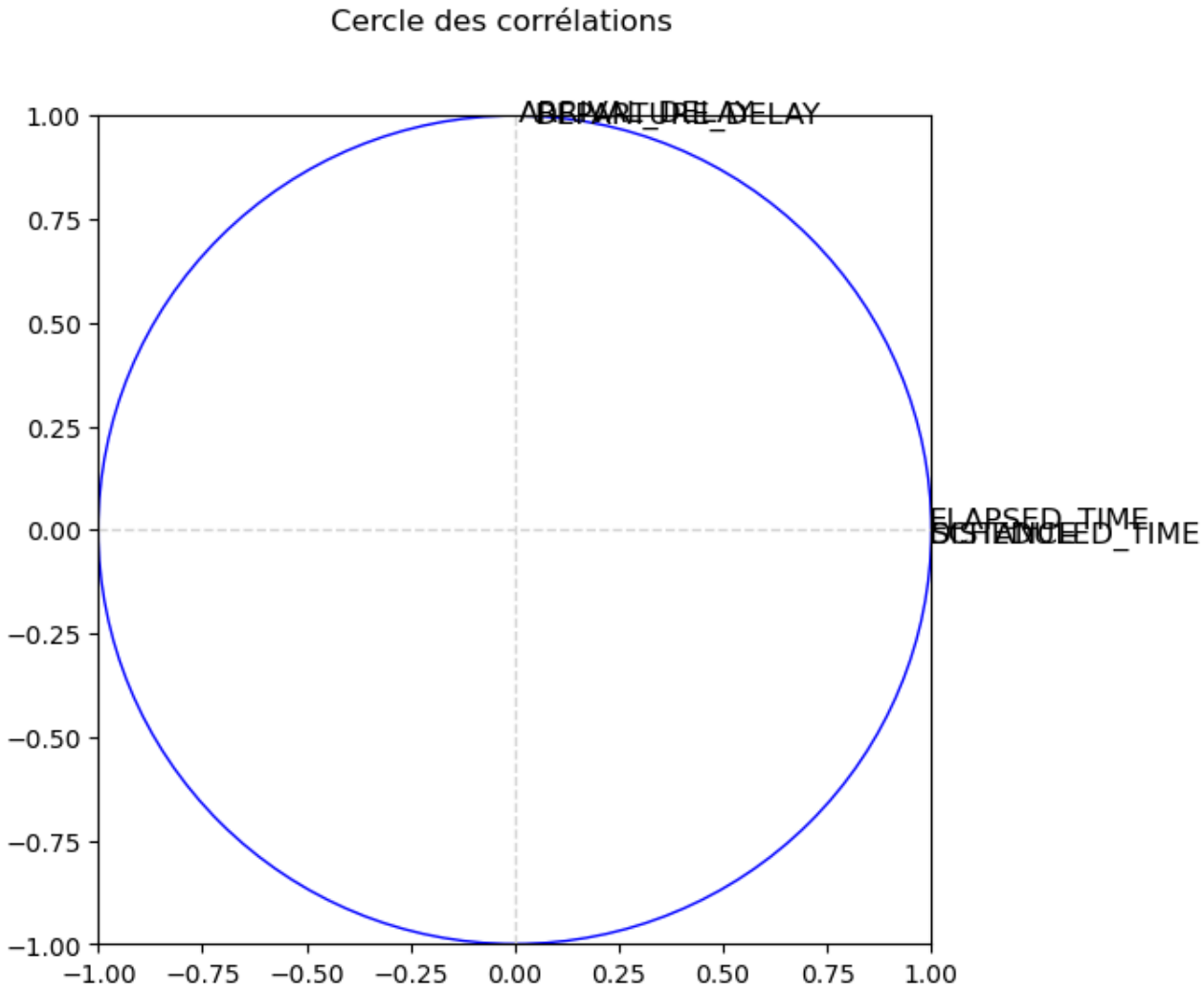


FIGURE 2.28 – Cercle de corrélation

Chapitre 3

Prédiction

Cette partie du rapport détaille le travail de prédiction réalisé.
Le but est de prédire le retard moyen que commettrait une compagnie aérienne pour un vol.

1. **Importation des bilbiothèques et pré-traitement du dataset**
   Vu la taille du dataset, il serait impossible de tester avec nos machines les modèles (cela prendrait trop de temps). Nous nous concentrons donc que sur le mois de janvier. Nous divisons le dataset en train et en test. Nous entraînons donc le modèle sur les trois premières semaines de janvier (échantillon train) et nous les testons sur les vols de la dernière semaine du mois (échantillon test). Afin de réaliser le travail, nous utilisons les bibliothèques **scikit-learn** d'où nous tirons les modèles.
2. **Les modèles**
   Nous testons 4 différents modèles :
   — Ridge Régression linéaire basique avec une norme L2 (pour la régularisation). Le but est de minimiser la fonction de coût suivante :

$$\sum (y_i - \hat{y}_i)^2 + \lambda \sum \beta_j^2$$

   — Lasso Régression linéaire basique avec une norme L1. Le but est de minimiser la fonction de coût suivante :

$$\sum (y_i - \hat{y}_i)^2 + \lambda \sum |\beta_j|$$

   — Elastic Net Combine les normes L1 et L2 pour bénéficier des avantages de Lasso et Ridge.

$$\sum (y_i - \hat{y}_i)^2 + \lambda_1 \sum |\beta_j| + \lambda_2 \sum \beta_j^2$$

   — RandomForest Algorithme d'ensemble qui combine plusieurs arbres de décision pour améliorer la performance et réduire le surapprentissage.
3. **Les métriques**
   Etant donné qu'on a affaire à une tâche de régression donc prédiction d'un retard moyen, on utilise les métriques suivantes :
   — Mean Squared Error (MSE) :
   Qui calcule la moyenne des carrés des erreurs.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

avec

- n : le nombre d'exemples
- y_i : la valeur de vérité terrain
- $\hat{y}_i$: la prédiction

— Root Mean Squared Error (RMSE) :
Qui calcule la racine carrée de la moyenne des carrées des erreurs, c'est une valeur plus facile à interpréter car elle est à la même échelle que les données. Dans notre cas elle représente l'écart en minutes entre les prédictions et les valeurs réelles.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Pour chaque modèle, on obtient les résultats suivants :

	Ridge	**Lasso**	**Elastic Net**	**Random Forest**
MSE	68.91	**66.30**	66.45	69.18
RMSE	8.30	**8.14**	8.15	8.32

TABLE 3.1 – Résultats de la prédiction sur les données test

On remarque que c'est le modèle Lasso qui présente les meilleures résultats.

4. **Test de significativité**
Toutefois, pour bien comparer les performances des modèles et vérifier si les différences sont significativement différentes, on applique un **test de significativité** en calculant la p-value. Si elle est inférieure à 0.05, alors les résultats sont significativement différents, sinon ils ne le sont pas. Dans notre cas, en comparant tous les modèles entre eux, les p-values étaient égales à 1 il n'y avait donc aucune différence significative de MSE ou de RMSE. Les quatres modèles performent au final de la même manière et présentent des résultats satisfaisants.

Conclusion

En conclusion, notre analyse révèle plusieurs tendances significatives dans l'industrie aéronautique. Les compagnies low-cost, malgré leur fréquence de vols élevée, utilisent moins d'avions long-courrier et desservent moins de destinations internationales comparées aux compagnies full-service. Toutefois, elles parviennent à compenser par une plus grande fréquence de vols, ce qui pourrait représenter un modèle économique avantageux en termes de coûts opérationnels. Les retards sont également plus fréquents chez les compagnies low-cost, bien que cela soit généralement compensé par une meilleure gestion des horaires.

L'analyse des avis clients a permis de mettre en évidence les facteurs clés de satisfaction et d'insatisfaction, fournissant des axes d'amélioration pour les compagnies aériennes et les aéroports. Ces résultats peuvent guider les stratégies futures pour l'ouverture et la gestion efficace d'une nouvelle compagnie aérienne sur le marché des États-Unis. De plus, le développement et l'application des modèles de machine learning ont démontré leur valeur pour la prédiction des retards.

Ce projet nous a également permis d'acquérir des compétences en scraping de données. Ces nouvelles compétences, ainsi que les outils que nous avons développés, fournissent désormais une base solide pour explorer de nouveaux domaines avec confiance.