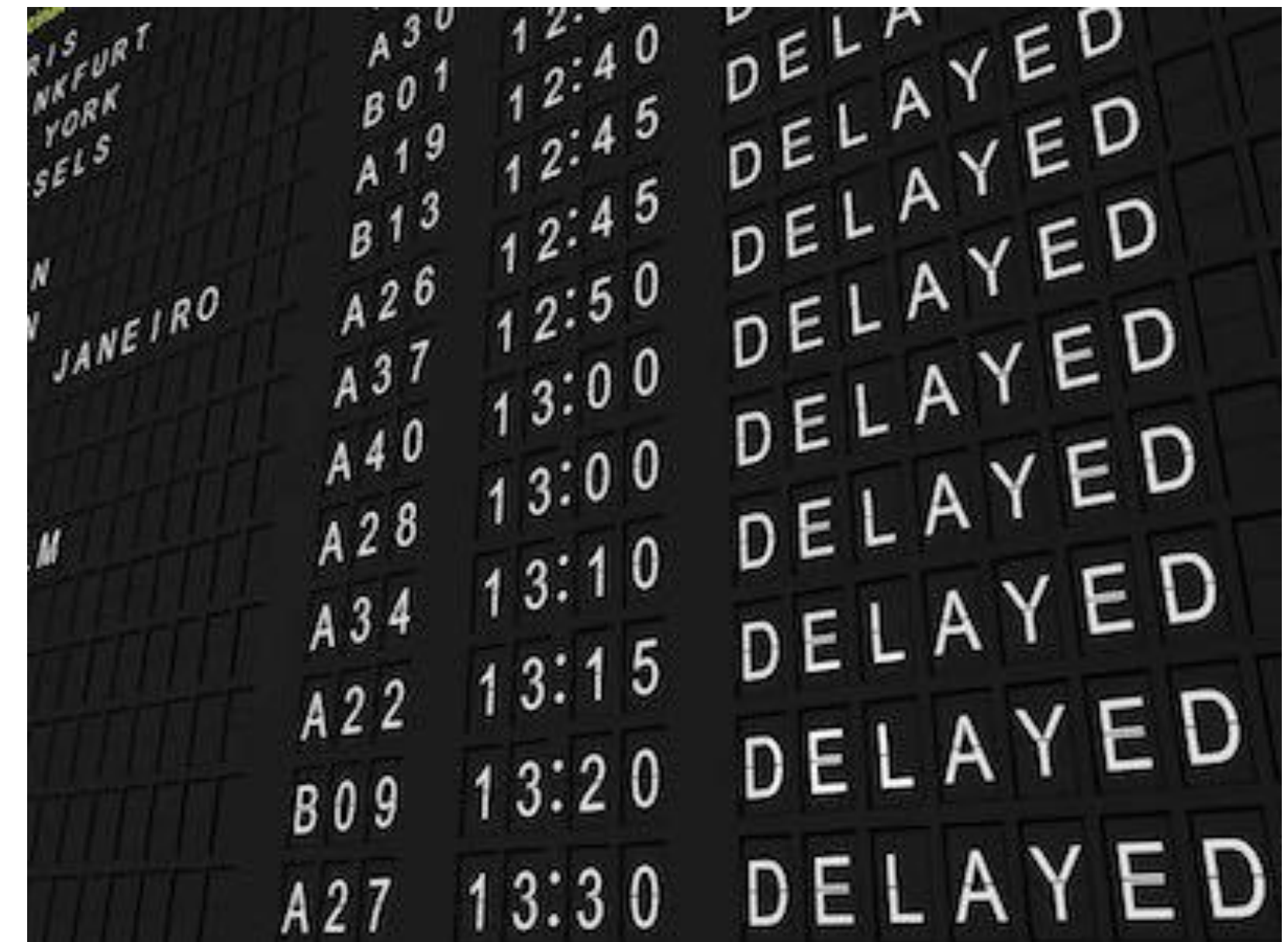


# ANTICIPEZ LE RETARD DE VOL DES AVIONS

Projet 4

Azim Makboulhoussen  
23 Février 2018



# Sommaire

- ❑ Introduction
- ❑ Analyse exploratoire des données
- ❑ Pistes de modélisations effectuées
- ❑ Choix du modèle finale et implémentation interface WEB
- ❑ Conclusion

# Introduction

# Objectif du projet

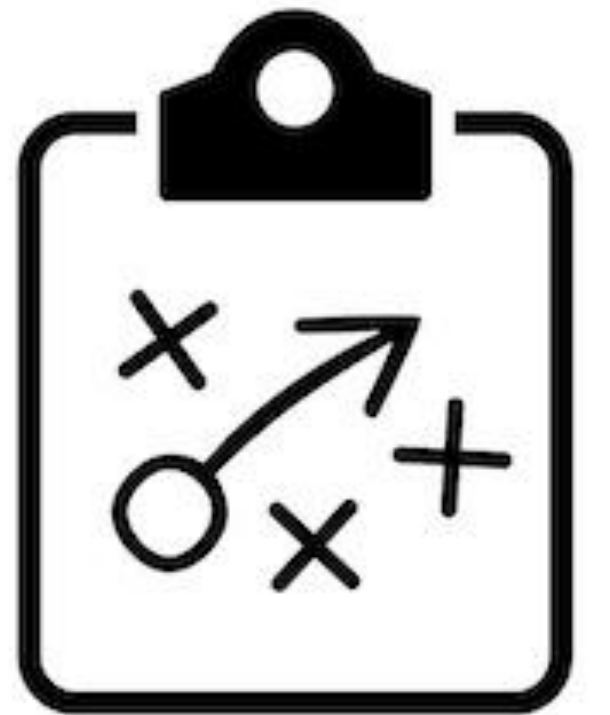


- ❑ Implémenter un modèle de prédiction des retards d'avions.
- ❑ Réaliser une analyse exploratoire de données d'informations sur des vols de différentes compagnies aériennes.
- ❑ Essayer différents modèles d'apprentissage machine afin de prédire le retard d'un vol
- ❑ Evaluation et amélioration des performances des modèles
- ❑ Implémentation d'un site WEB utilisant le modèle choisi de prédiction des retards.

# Exploration des données

# Stratégie d'exploration des données

- ❑ Chargement d'un seul fichier (1 mois) en raison de la volumétrie des données
- ❑ Analyse rapide des données
- ❑ Observation des valeurs manquantes
- ❑ Choix des variables pertinentes
- ❑ Chargement de l'ensemble des données en se basant sur les variables sélectionnées
- ❑ Analyse exploratoire des données sur ces données



# Les données



- ❑ Information sur les vols d'avions de 2016 contenu dans 12 fichiers : 1 fichier par mois
- ❑ Au total, les données représentent plus de **5 millions** de vols (**environ 2 Gb**)
- ❑ Chaque vol est décrit par **65** variables

AR	QUARTER	MONTH	DAY_OF_MONTH	DAY_OF_WEEK	FL_DATE	UNIQUE_CARRIER	AIRLINE_ID	CARRIER	TAIL_NUM	FL_NUM	ORIGIN_AIRPORT_ID	ORIGIN
2016	2	6	4	6	2016-06-04	DL	19790	DL	N908DL	1138	10397	
2016	2	6	4	6	2016-06-04	DL	19790	DL	N924DN	1139	11278	
2016	2	6	4	6	2016-06-04	DL	19790	DL	N904DE	1140	10397	
2016	2	6	4	6	2016-06-04	DL	19790	DL	N817DN	1143	14122	
2016	2	6	4	6	2016-06-04	DL	19790	DL	N929DL	1145	10397	



# Caractéristiques des données

Feature	Description
YEAR	Année du vol
QUARTER	Trimestre (1 à 4)
MONTH	Mois : 1 à 12
DAY_OF_MONTH	Le jour
DAY_OF_WEEK	Lundi à Dimanche (1 à 7)
FL_DATE	Date du vol (yyyymmdd)
UNIQUE_CARRIER	Code unique identifiant la compagnie
AIRLINE_ID	Identifiant unique assigné par US DOT pour une compagnie. On remarque une valeur anormale
CARRIER	Code IATA pour identifier une compagnie. Le code n'est pas toujours unique. Il vaut mieux utiliser Unique Carrier
TAIL_NUM	Immatriculation de l'avion
FL_NUM	Numéro de vol
ORIGIN_AIRPORT_ID	Identifiant par US DOT pour identifier un aéroport de manière unique
ORIGIN_AIRPORT_SEQ_ID	Identifiant aéroport mais à un moment donné
ORIGIN_CITY_MARKET_ID	Identifiant de la ville (city market)
ORIGIN	Code de l'aéroport d'origine
ORIGIN_CITY_NAME	Ville de l'aéroport d'origine
ORIGIN_STATE_ABR	Etat d'appartenance de l'aéroport
ORIGIN_STATE_FIPS	Etat d'appartenance de l'aéroport
ORIGIN_STATE_NM	Etat d'appartenance de l'aéroport
ORIGIN_WAC	Aéroport d'origine, World Area Code
DEST_AIRPORT_ID	Identifiant par US DOT pour identifier un aéroport de manière unique
DEST_AIRPORT_SEQ_ID	Identifiant aéroport mais à un moment donné
DEST_CITY_MARKET_ID	Identifiant de la ville (city market)
DEST	Code de l'aéroport d'origine
DEST_CITY_NAME	Ville de l'aéroport d'origine
DEST_STATE_ABR	Etat d'appartenance de l'aéroport

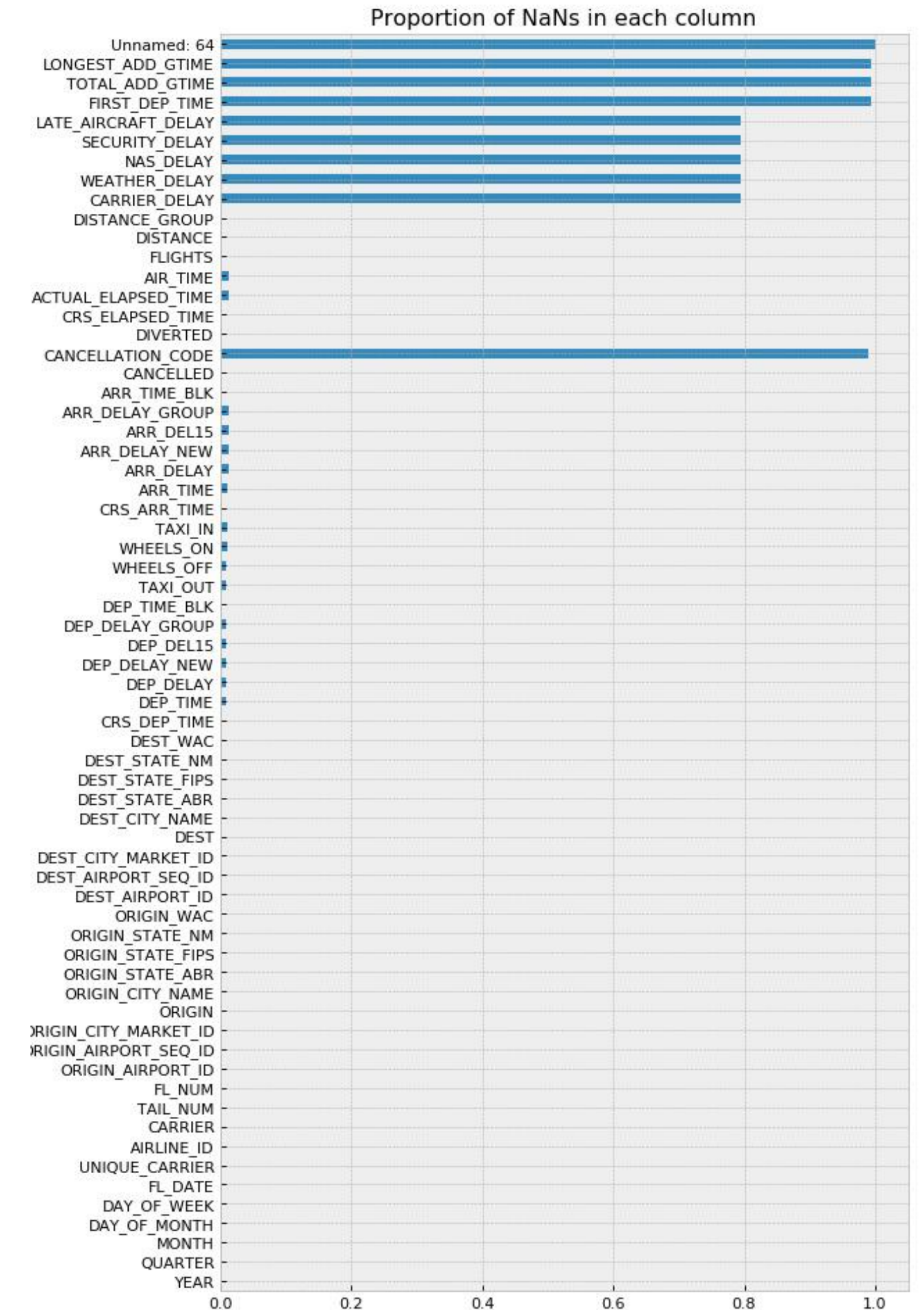
- ❑ Compréhension de chacune des variables
- ❑ Les données sont composées de variables :
  - **catégorielles** comme le code de la compagnie, le code de l'aéroport origine et destination, ...
  - **continues** comme la distance, le retard à l'arrivée ou au départ, le nombre de vols, ...

DEST_STATE_FIPS	Etat d'appartenance de l'aéroport
DEST_STATE_NM	Etat d'appartenance de l'aéroport
DEST_WAC	Aéroport d'origine, World Area Code
CRS_DEP_TIME	Heure de départ prévu(hhmm)
DEP_TIME	Heure de départ réel (hhmm)
DEP_DELAY	Différence entre schedule et actual time en minute
DEP_DELAY_NEW	Même chose que dep_delay mais pas de valeur négative
DEP_DEL15	1 = retard ou pas
DEP_DELAY_GROUP	code qui indique la valeur du retard (- 15 min, entre 15 et 29 min, ...)
DEP_TIME_BLK	Heure de départ prévu (dans un groupe)
TAXI_OUT	Roulage sur la piste (en minute)
WHEELS_OFF	Roue rentré en hh:mi
WHEELS_ON	Roue sorti en hh:mi
TAXI_IN	Roulage sur la piste à l'arrivée
CRS_ARR_TIME	Heure d'arrivée prévue
ARR_TIME	Heure d'arrivée réelle
ARR_DELAY	Différence entre schedule et actual time en minute
ARR_DELAY_NEW	Même chose que dep_delay mais pas de valeur négative
ARR_DEL15	1 = retard ou pas
ARR_DELAY_GROUP	code qui indique la valeur du retard (- 15 min, entre 15 et 29 min, ...)
ARR_TIME_BLK	Heure d'arrivée prévu (dans un groupe)
CANCELLED	1 = vol annulé
CANCELLATION_CODE	code qui indique la raison
DIVERTED	1 = dévié
CRS_ELAPSED_TIME	Durée du vol prévue
ACTUAL_ELAPSED_TIME	Durée du vol réel
AIR_TIME	durée du vol entre le moment où il quitte le sol et il touche le sol
FLIGHTS	Nb de vols
DISTANCE	Distance en miles
DISTANCE_GROUP	code pour interval de distance
CARRIER_DELAY	retard compagnie
WEATHER_DELAY	retard météo
NAS_DELAY	retard NAS
SECURITY_DELAY	retard sécurité
LATE_AIRCRAFT_DELAY	retard avion
FIRST_DEP_TIME	1er gate heure de départ
TOTAL_ADD_GTIME	
LONGEST_ADD_GTIME	



# Valeurs manquantes

- ❑ Certaines colonnes contiennent énormément de valeurs vides (LATE\_AIRCRAFT\_DELAY, NAS\_DELAY, CARRIER\_DELAY, ...)
- ❑ La plupart des colonnes est bien remplie.
- ❑ Les colonnes contenant beaucoup de valeurs vides (80%) ne seront pas conservées.



# Sélection de variables pertinentes

MONTH	Le mois est important car on peut avoir des retards plus importants sur certains mois.
DAY_OF_MONTH	Idem, certains jours du mois peuvent connaître plus de vols avec retard
DAY_OF_WEEK	Plus de vols certains jours de semaine, donc potentiellement plus de retard.
UNIQUE_CARRIER	Permet d'identifier la compagnie.
FL_NUM	Numéro de vol. Certain vols peuvent connaître régulièrement des retards
ORIGIN	Code de l'aéroport de départ. Important car il peut y avoir plus de problèmes (organisations, ...) dans certains aéroports qui peuvent connaître plus de fréquence de retards.
DEST	Même raisonnement que la variable ORIGIN
CRS_DEP_TIME	Heure de départ prévue. Peut influencer sur le retard.
CRS_ARR_TIME	Heure d'arrivée prévue. Peut influencer sur le retard.
DEP_DELAY	Retard au départ. Important pour notre modèle car va jouer sur le retard global d'un vol.
ARR_DELAY	C'est notre target variable. Celle qu'on devra prédire.
DISTANCE	la distance peut influencer aussi sur le retard
CRS_ELAPSED_TIME	Durée de vol prévue. La durée peut aussi influencer les retards.
ACTUAL_ELAPSED_TIME	Durée de vol réel. Influe directement sur le retard potentiel.
DIVERTED	indique si le vol a été dévié. Un vol dévié aura forcément du retard.
CANCELLED	si le vol est souvent annulé, il a des chances d'avoir souvent des retards
FL_DATE	information sur la date du vol

# Traitement sur les données

- ❑ **Valeurs manquantes :**

- Suppression des lignes contenant des valeurs vides (% très faible par rapport à la quantité de données)

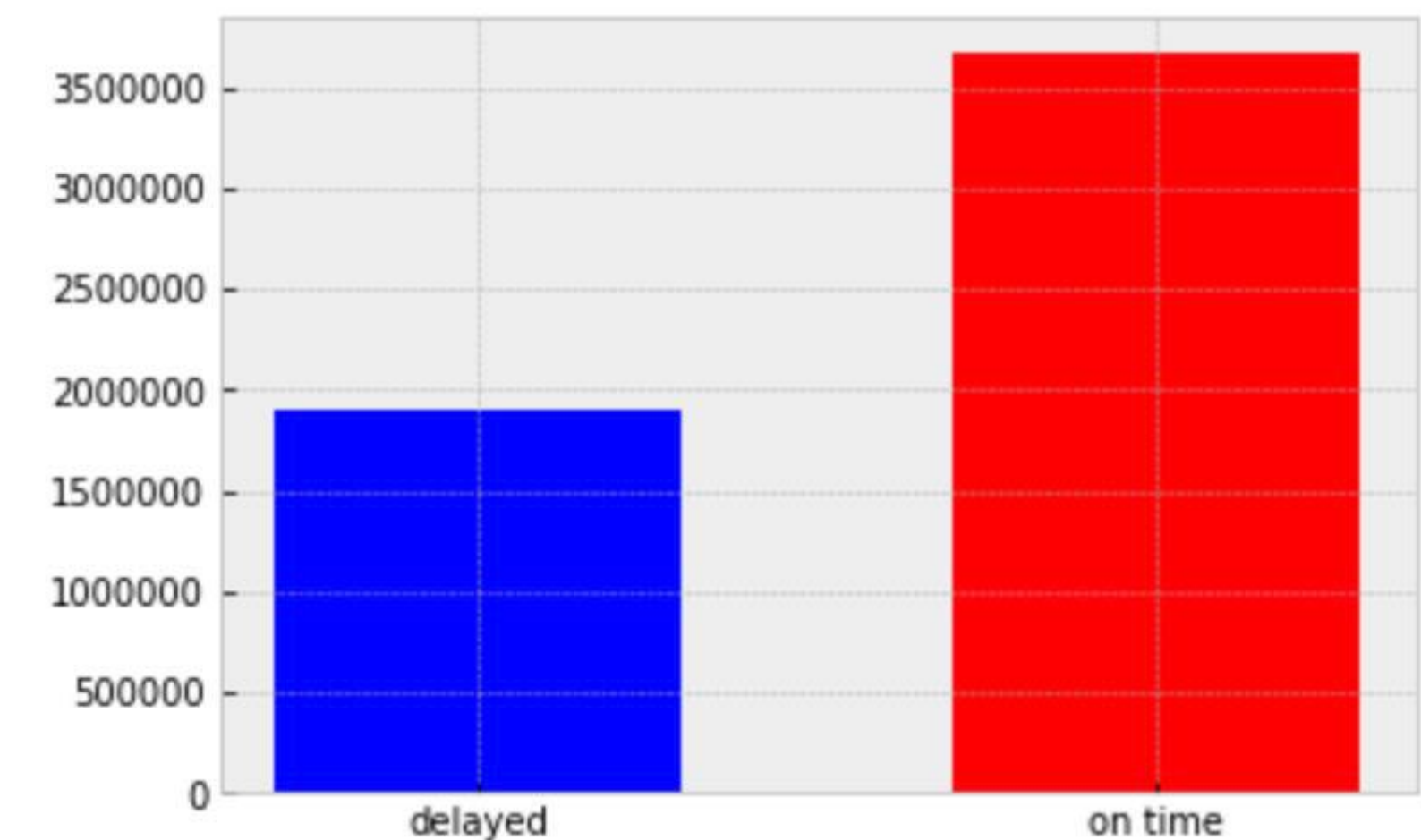
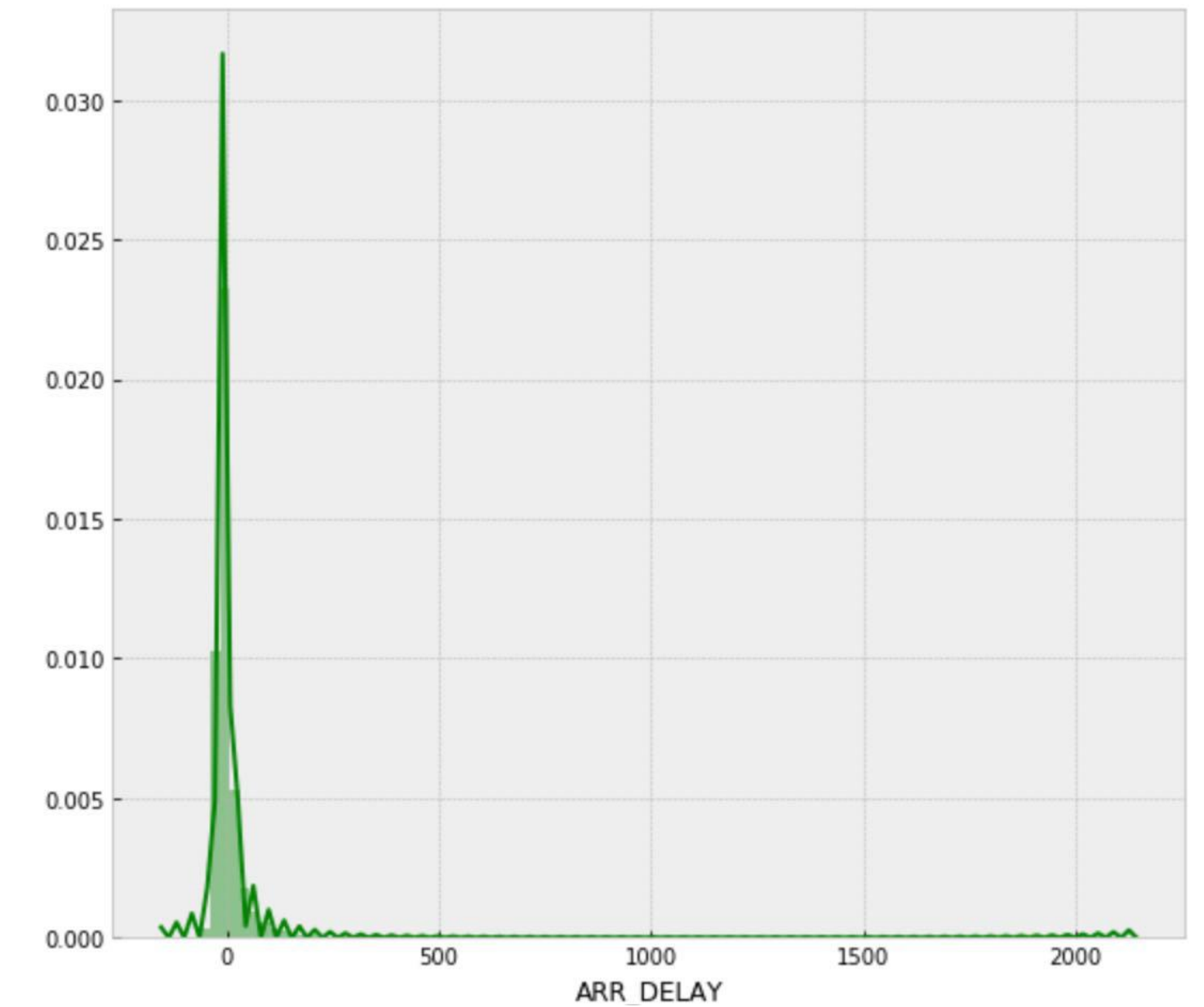
- ❑ Suppression des variables DIVERTED et CANCELED qui sont toujours à 0.

- ❑ **Feature Engineering :**

- colonne contenant le **numéro de semaine**
- colonne contenant **la proximité à un jour férié** (en nombre de jours)
- **Heure de départ** et **heure d'arrivée** (sans les minutes)
- Code numérique pour le **code compagnie**
- Code numérique pour le **code aéroport**

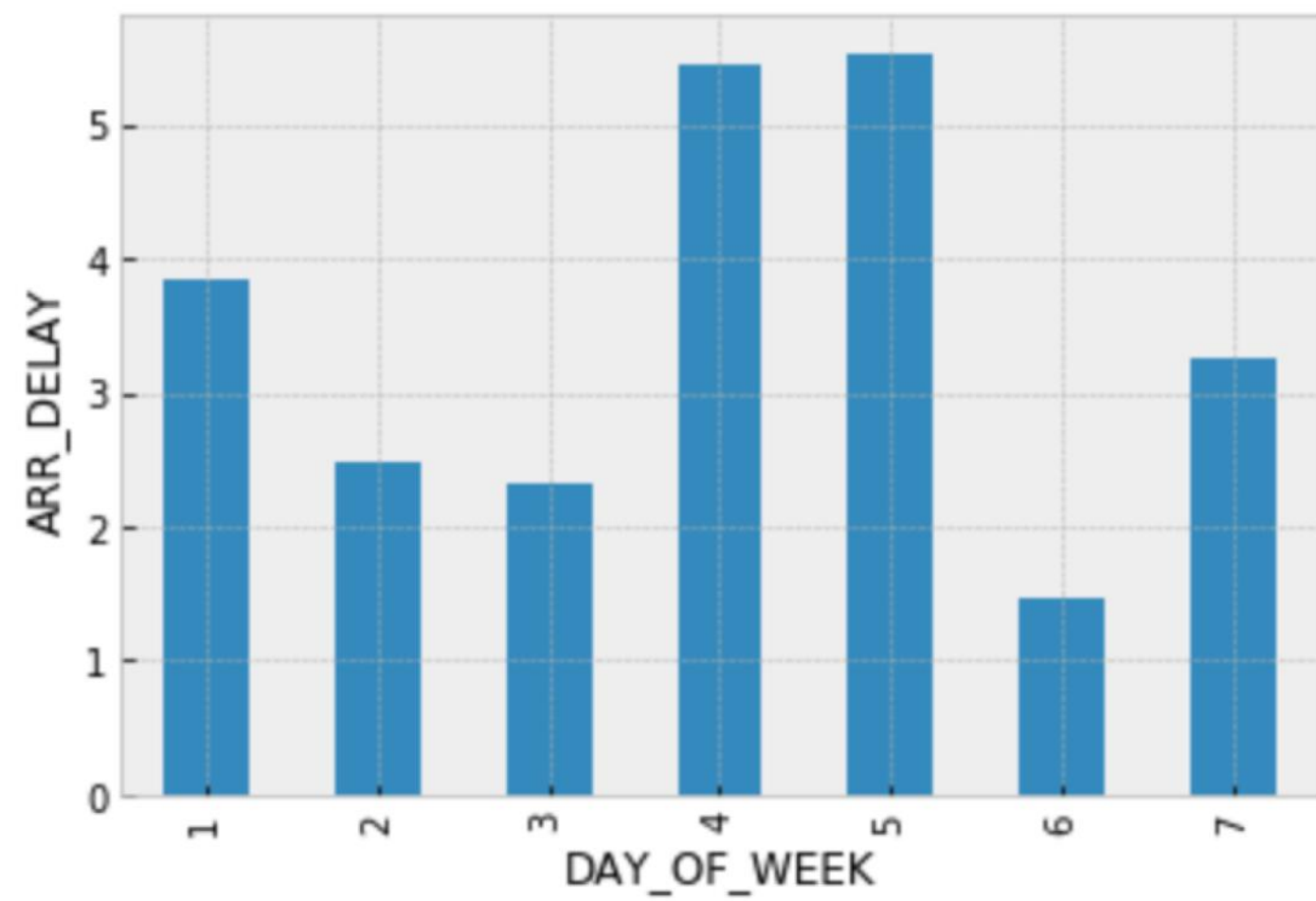
# Analyse de la variable cible

- ❑ Notre variable cible est : **ARR\_DELAY**
- ❑ La distribution est centrée autour de 0 ce qui signifie que globalement les avions sont à l'heure ou ils ont peu de retard ou d'avance.
- ❑ Seuls 33% des vols sont en retard.
- ❑ La moyenne des retards est de 3 min (et la médiane : 6 min d'avance).
- ❑ Nombre important de valeurs extrêmes.

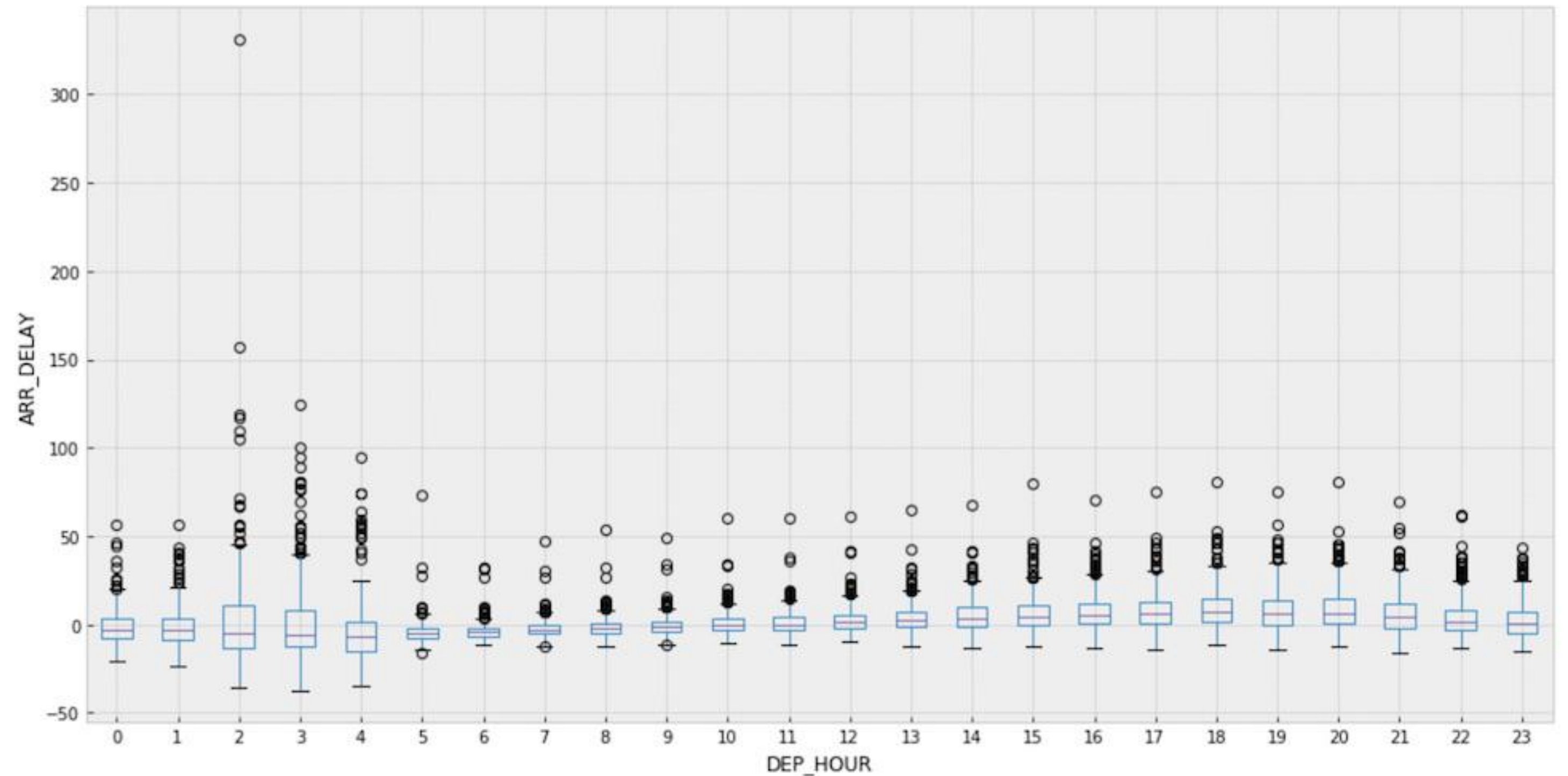




# Analyse des retards

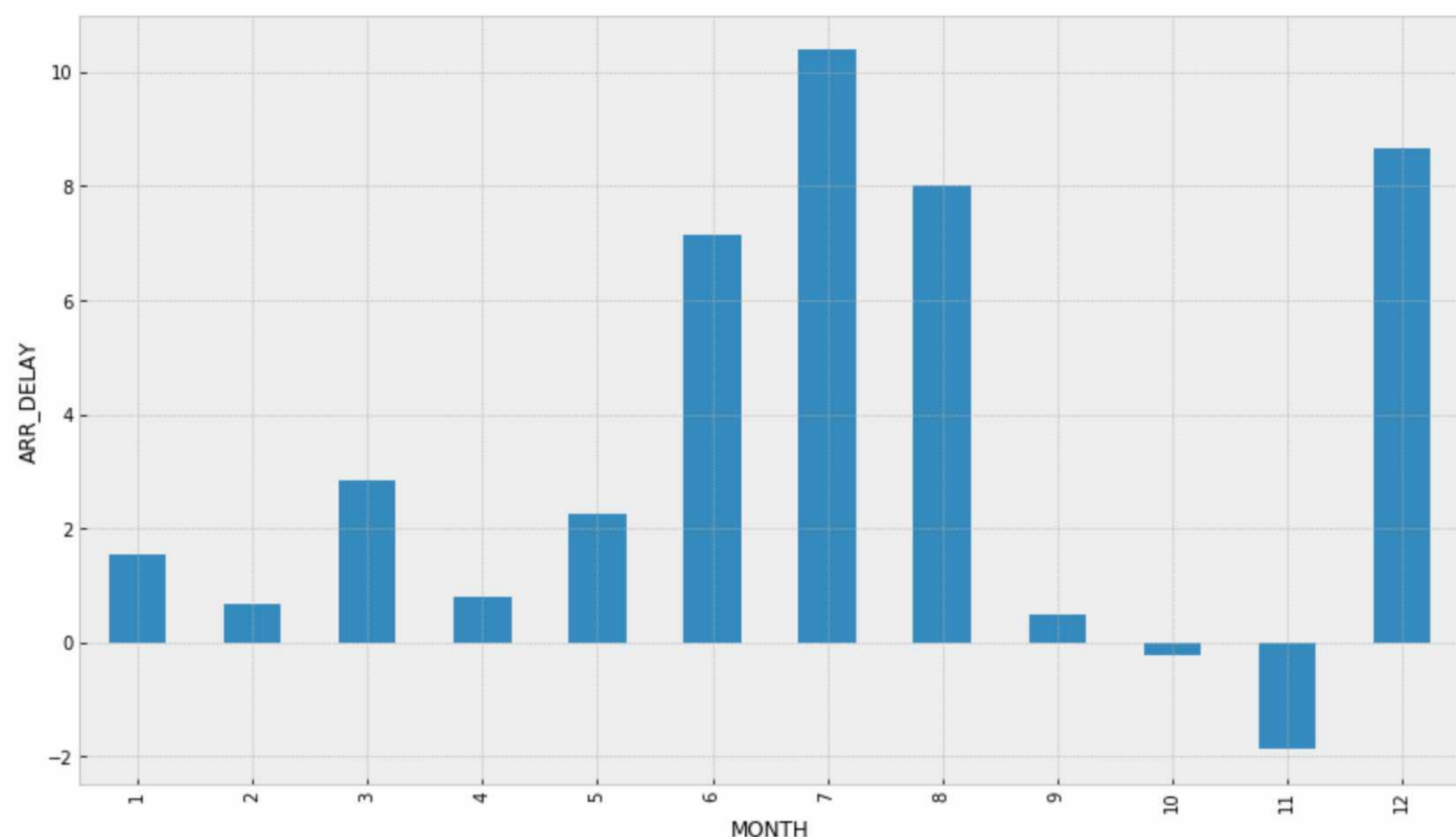


- ❑ Nous avons plus de retards le jeudi et vendredi
- ❑ Samedi meilleur jour pour arriver à l'heure

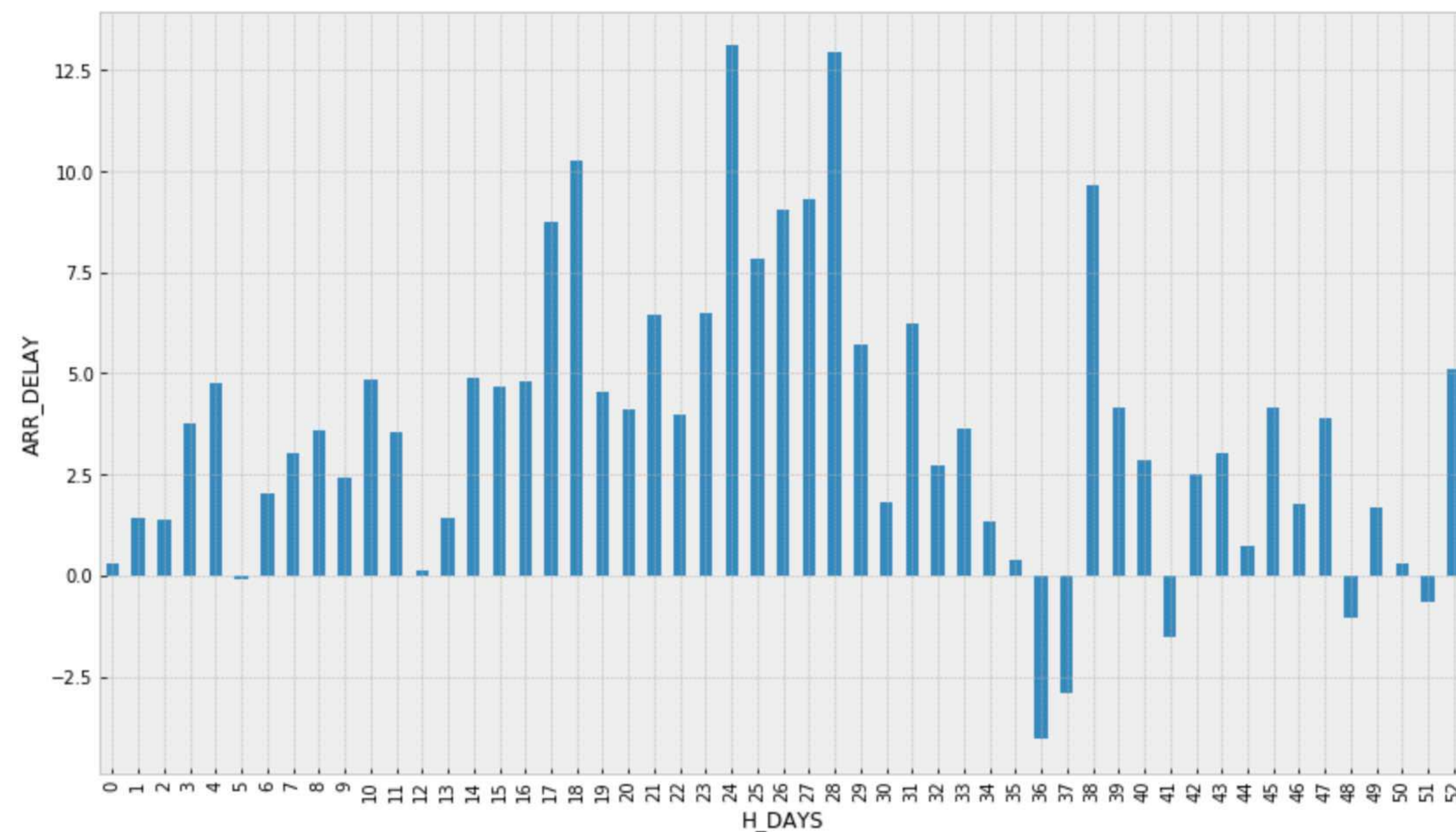


- ❑ Nous avons beaucoup moins de retard le matin
- ❑ Il vaut mieux éviter les horaires entre 14h et 22h pour ne pas arriver en retard
- ❑ Avant 7h les vols sont plus réguliers en terme de ponctualité

# Analyse des retards



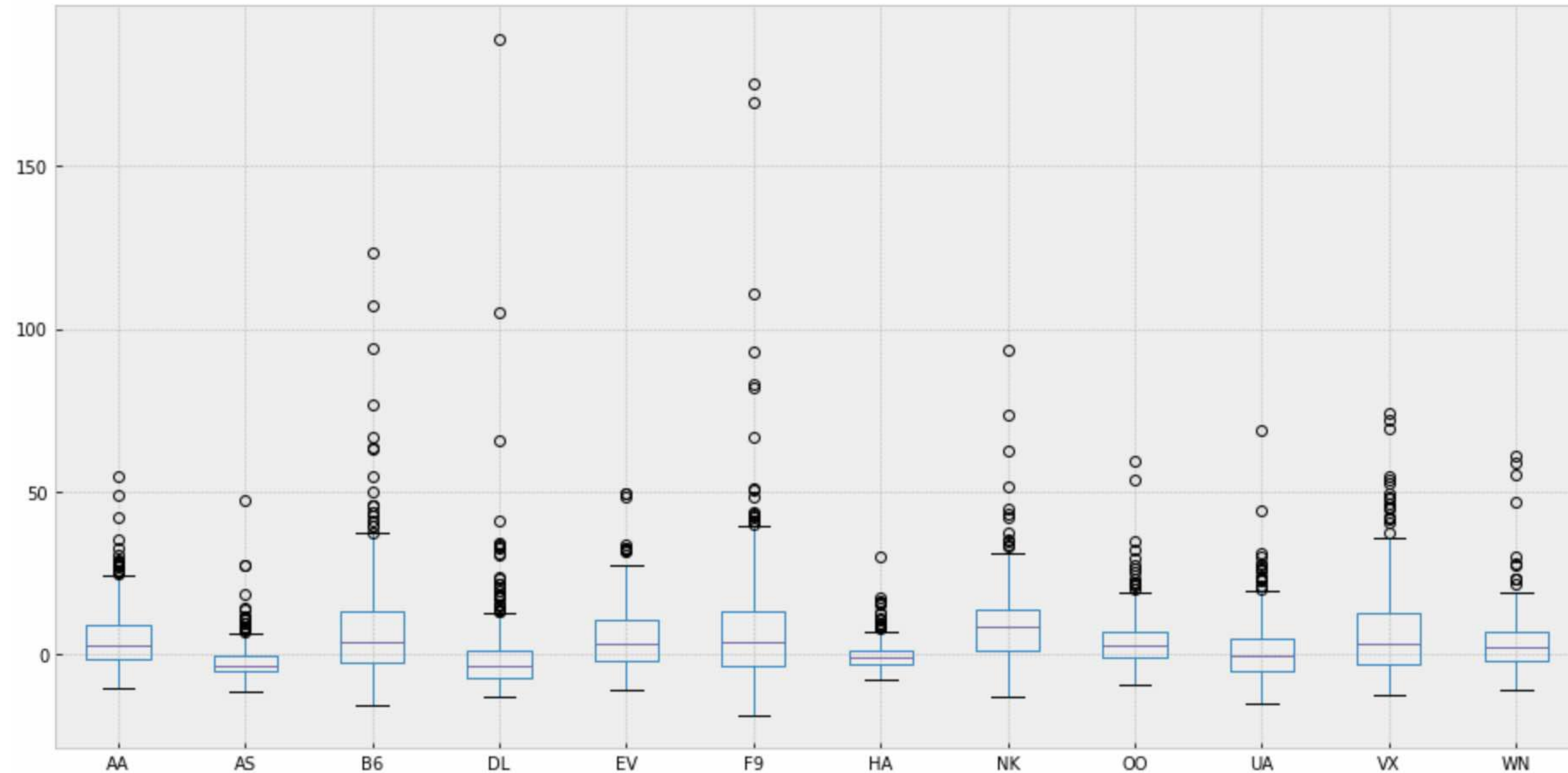
- ❑ Nous avons plus de retard le Juin, Juillet, Août et Décembre
- ❑ Le mois de Novembre est celui où on a plus de chance d'arriver à l'heure.



- ❑ C'est surtout entre 15j et 30j autour des jours fériés que nous avons des retards.

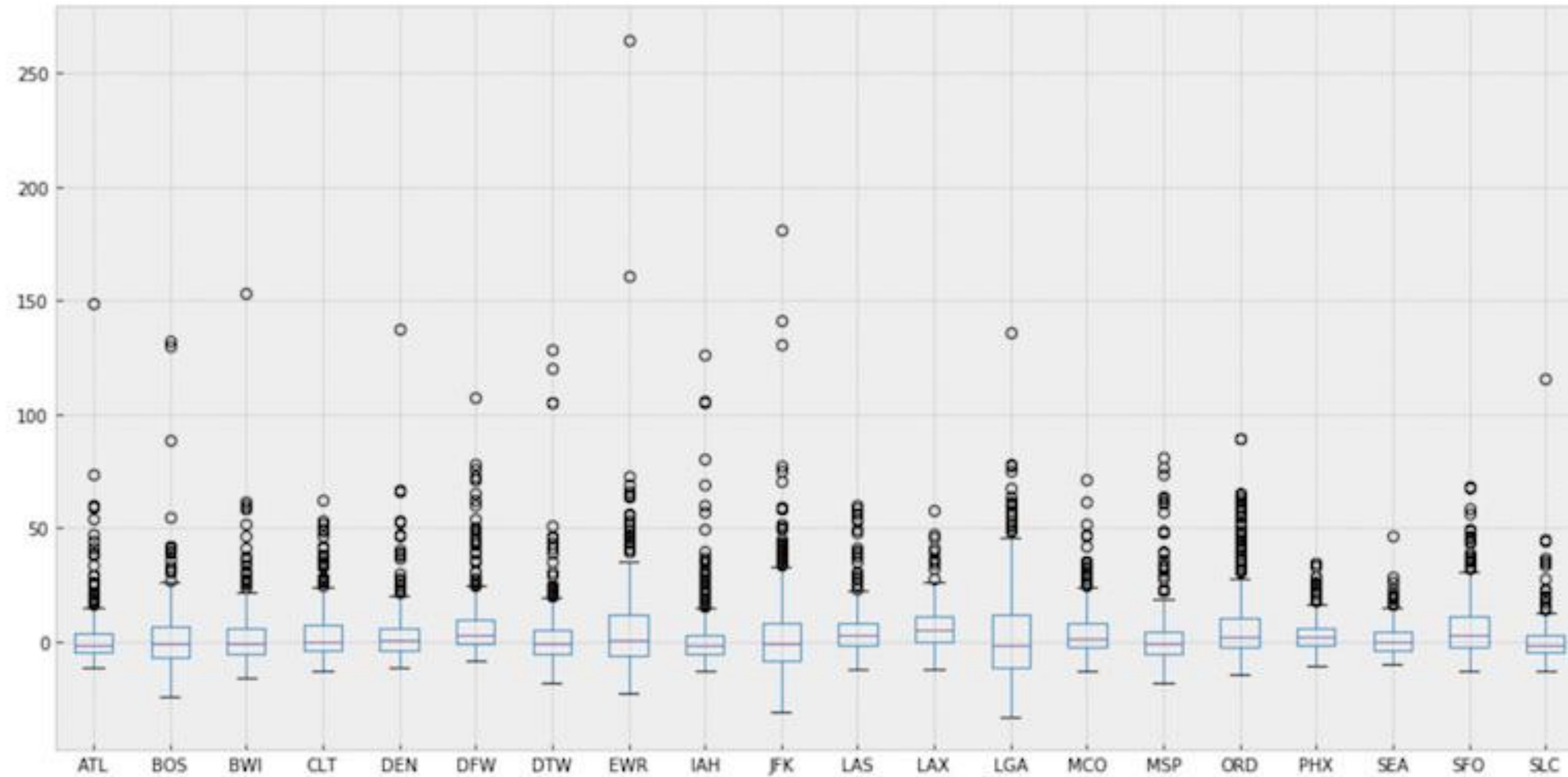


# Analyse des retards - compagnies



- ❑ Alaska Airlines (AS), Delta Airlines (DL) et Hawaï Airlines (HA) sont plus fiables en terme de ponctualité
- ❑ Frontier Lines (F9), South West (NK) et Jet Blue (B6) sont plus en retard en moyenne
- ❑ Nous avons des valeurs extrêmes pour Frontier Lines et Delta Airlines

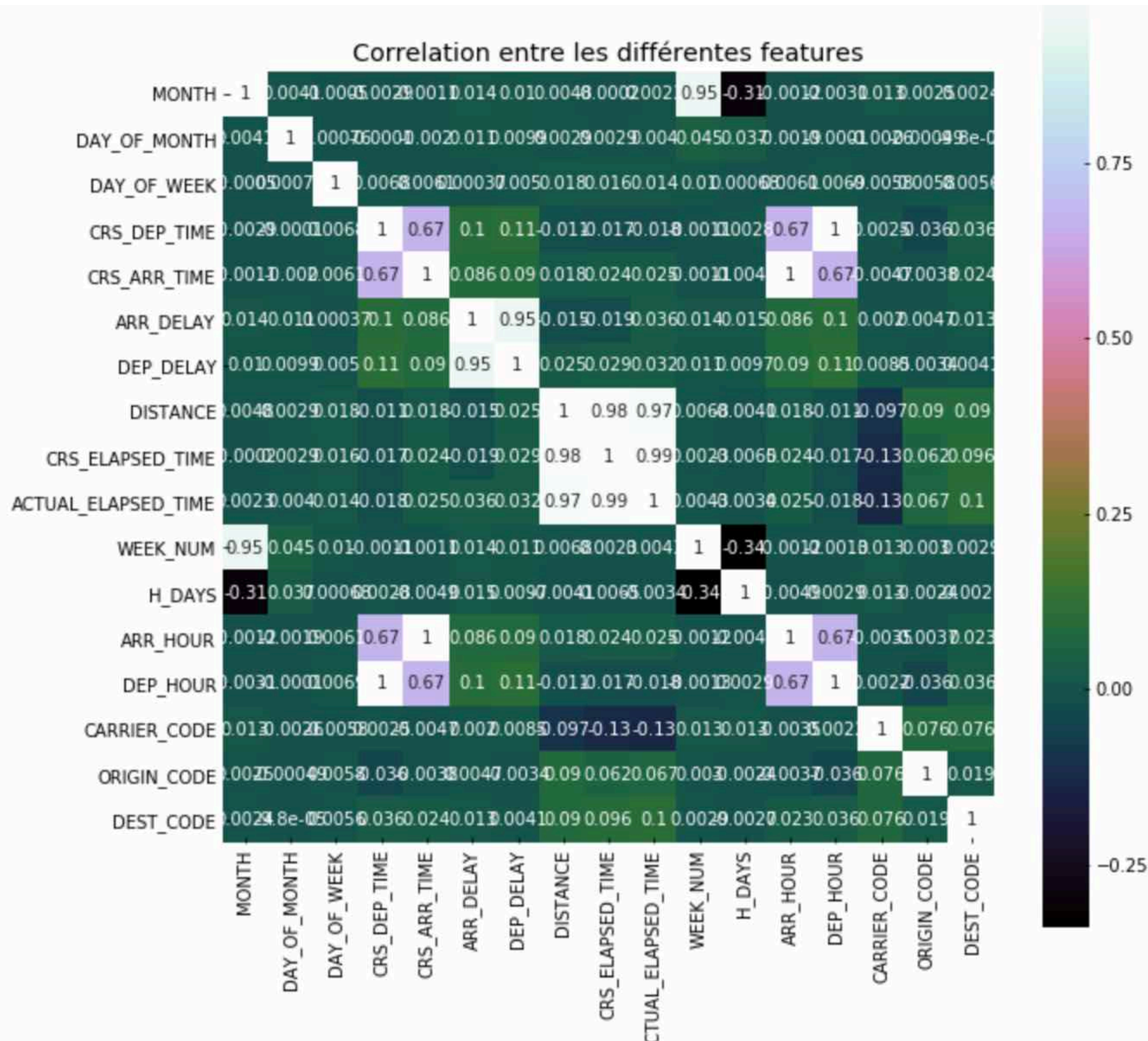
# Analyse des retards - aéroports



- ❑ Les aéroports de Dallas (DFW), Las Vegas (LAX) ou Los Angeles (LGA) semblent plus fiables en terme de ponctualité que ceux de de Salt Lake City (SLC), Houston (IAH) ou Atlanta (ATL).



# Corrélations des variables

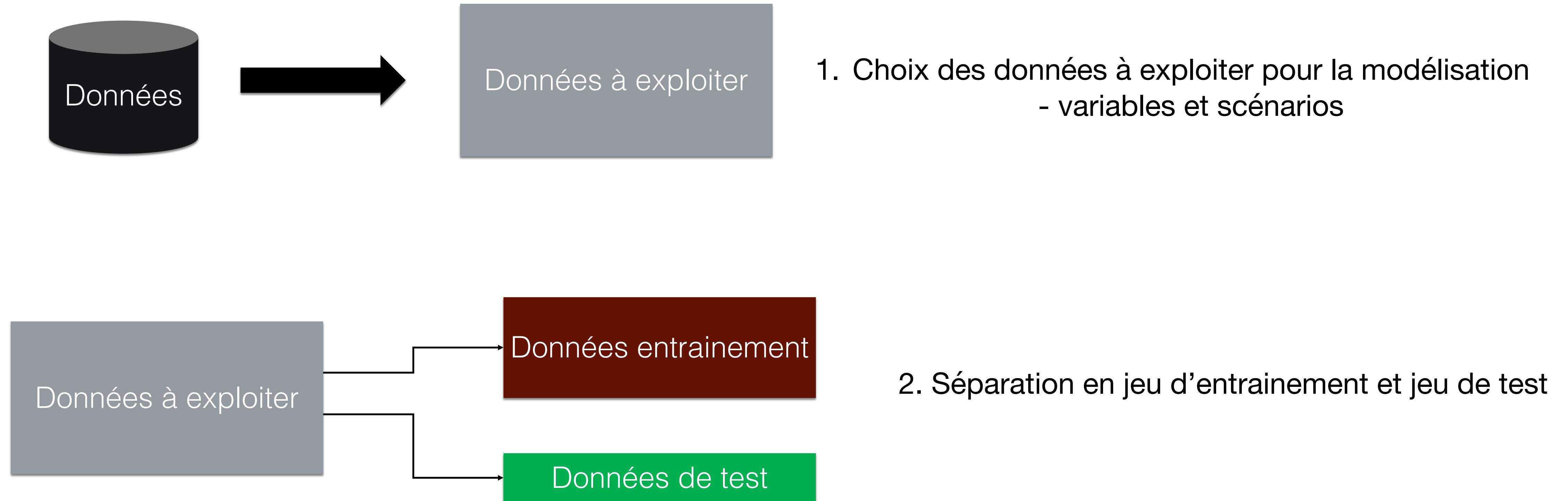


## Corrélations →

- ❑ **CRS\_DEP\_TIME** et **CRS\_ARR\_TIME**
  - ❑ **ARR\_DELAY** et **DEP\_DELAY**
  - ❑ **MONTH** et **WEEK\_NUM**
  - ❑ **CRS\_ELAPSED\_TIME** et **DISTANCE**
  - ❑ **ACTUAL\_ELAPSED\_TIME** et **CRS\_ELAPSED\_TIME**
  - ❑ **DEP\_HOUR** et **ARR\_HOUR**
- 
- ❑ Suppression **CRS\_DEP\_TIME** et **CRS\_ARR\_TIME** (DEP\_HOUR et DEP\_TIME)
  - ❑ Suppression **DEP\_DELAY** => **ARR\_DELAY**
  - ❑ Suppression **ACTUAL\_ELAPSED\_TIME, CRS\_ELAPSED\_TIME** => **DISTANCE**
  - ❑ Corrélation **DEP\_HOUR** et **ARR\_HOUR**. Conservation **DEP\_HOUR** uniquement dans la perspective API.

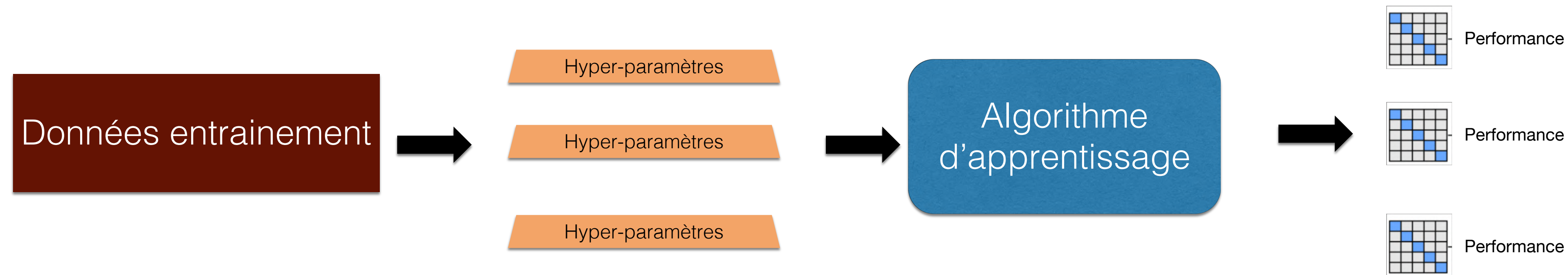
Pistes de modélisations

# Notre démarche d'évaluation de modèles





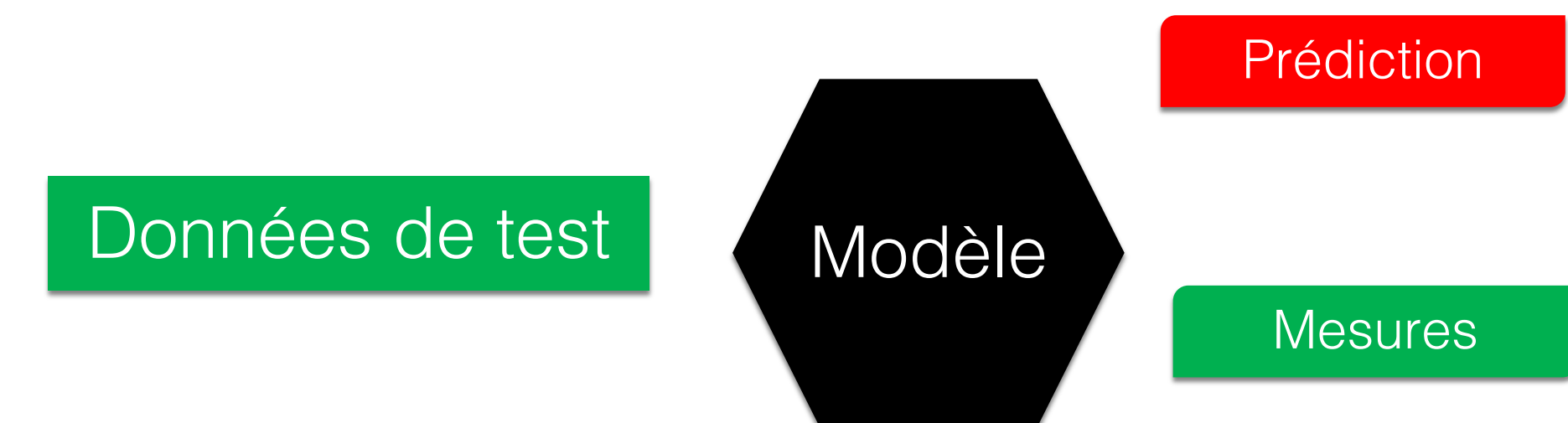
# Notre démarche d'évaluation de modèle



3. Evaluation de différentes valeurs d'hyper-paramètres par une recherche sur grille et une validation croisée pour trouver la meilleure performance



4. On applique alors les valeurs des paramètres les meilleurs aux données d'entraînement pour obtenir notre modèle



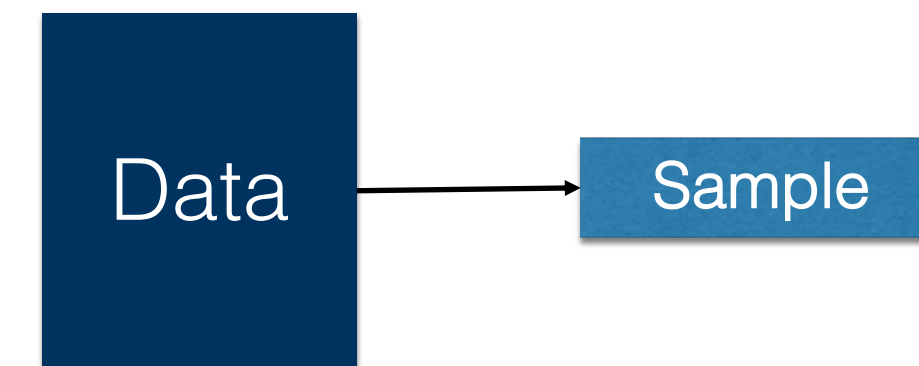
5. Evaluation du modèle avec les données de test



# Nos scénarii des tests des modèles

## ❑ Scénario 1 :

- Echantillon de 10% des données (500 000 vols)



## ❑ Scénario 2 :

- On filtre les données pour supprimer les vols en avance
- Echantillonnage à 500 000 vols



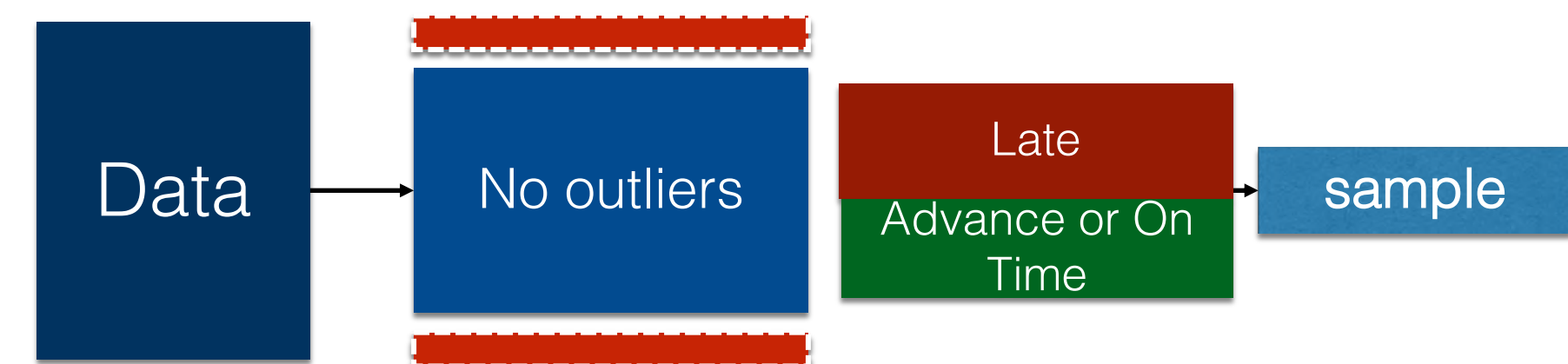
## ❑ Scénario 3 :

- Suppression des valeurs extrêmes de retards et avances Echantillon de 10%



## ❑ Scénario 4 :

- Suppression des valeurs extrêmes (retards et avances)
- Equilibrage entre nombre de retard et nombre d'avance
- Echantillonnage à 500 000 vols



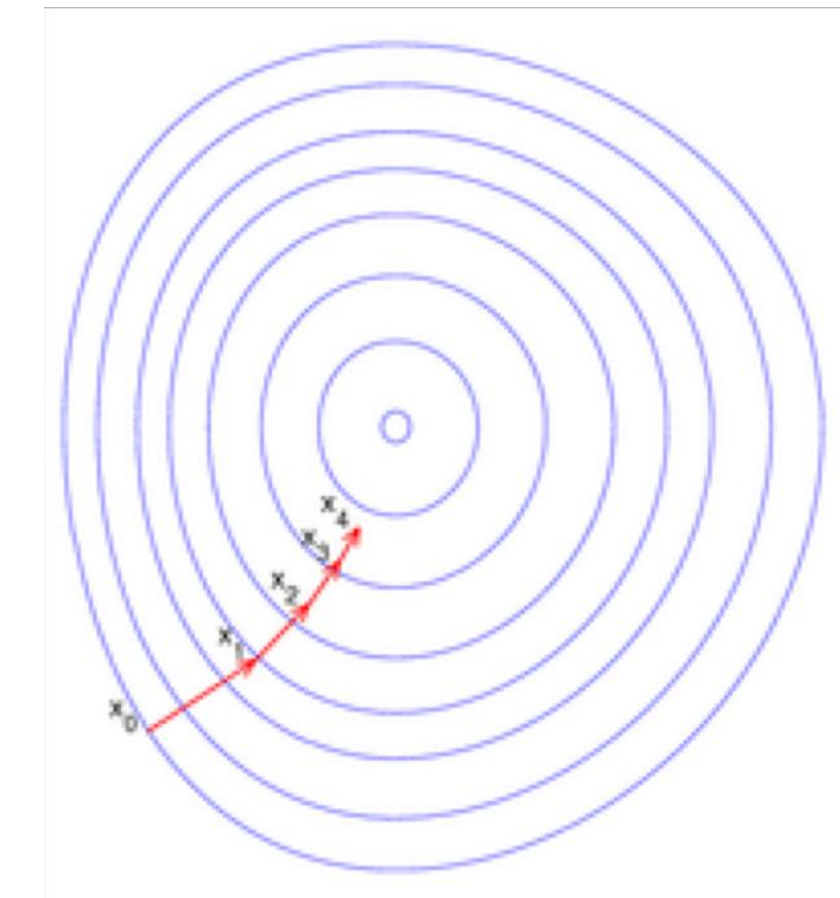
# Notre stratégie d'apprentissage

## □ Type d'apprentissage :

- Supervisé : le retard comme variable de réponse.
- Régression : la variable de réponse est continue.

## □ Choix algorithmes:

- Régression linéaire multiple pénalisée
- Contraste : fonction de **perte** + terme de **pénalité**



## □ Optimisation de l'algorithme :

- Utilisation de l'algorithme de **descente de gradient stochastique**
- Méthode itérative qui va chercher à minimiser la fonction de perte en optimisant les coefficients
- Adaptée pour de gros volume de données

# Notre stratégie d'amélioration de la performance

## ❑ Régularisation (pénalité):

- Ridge : restriction amplitude des variables (pénalité norme l2)
- Lasso : réduit la dimension en estimant à 0 les effets peu important (l1)
- Elasticnet : combinaison des 2

## ❑ Fonctions de perte :

- Squared loss (valeur réelle – valeur prédite)<sup>2</sup>
- Huber loss : perte quadratique modifiée pour être moins sensible aux données extrêmes.

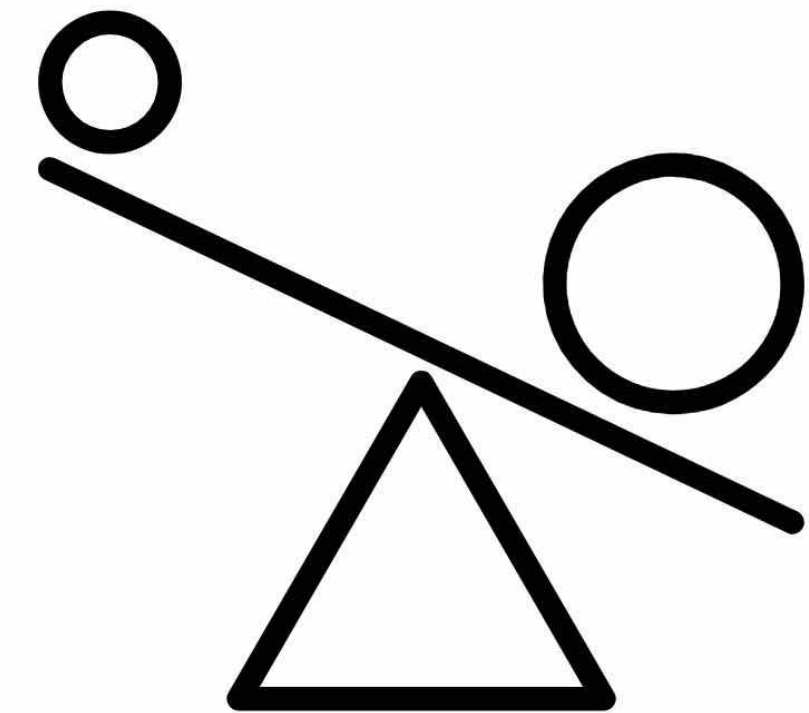
## ❑ Hyper paramètres :

- Alpha : coef de pénalité
- L1\_ratio : importance perte l1 dans la combinaison elasticnet

<b>loss</b>	squared_loss / huber
<b>penalty</b>	none / l1 / l2
<b>alpha</b>	[100 - 10 <sup>-7</sup> ]
<b>l1_ratio</b>	[.05, .15, .5, .7, .9, .95, .99, 1]

# Evaluation des algorithmes

- ❑ Validation croisée et recherche sur grille :
  - Donne les meilleurs paramètres au sens de la fonction de perte
  
- ❑ Comparaison entre les scénarii :
  - Utilisation de l'erreur quadratique moyenne (MSE) et l'erreur absolue moyenne (MAE) => la hiérarchie des modèles est souvent respectée par les 2 mesures.
  - Affichage des valeurs prédites vs valeurs de tests pour un échantillon (permet d'identifier ce que le modèle rate en terme de prédiction).
  - Affichage de la distribution des résidus



Résultats et implémentation

# Résultats

Pas de filtre

Uniquement retards

Sans extrêmes

Sans extrême / équilibré

MAE

21,08

31,13

12,24

10,90

Best Params

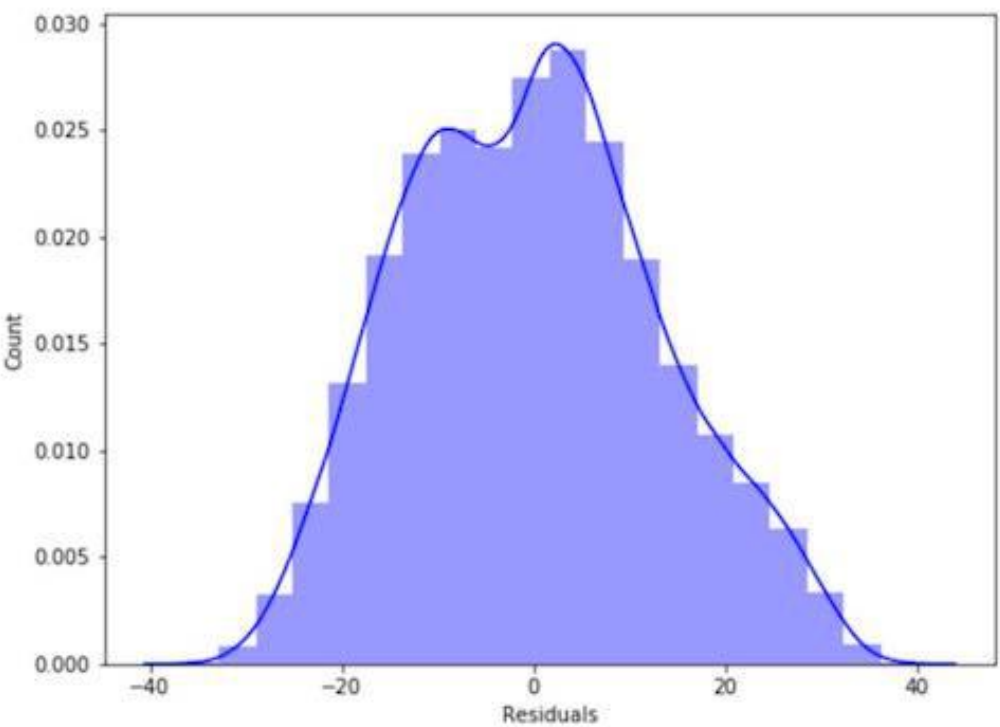
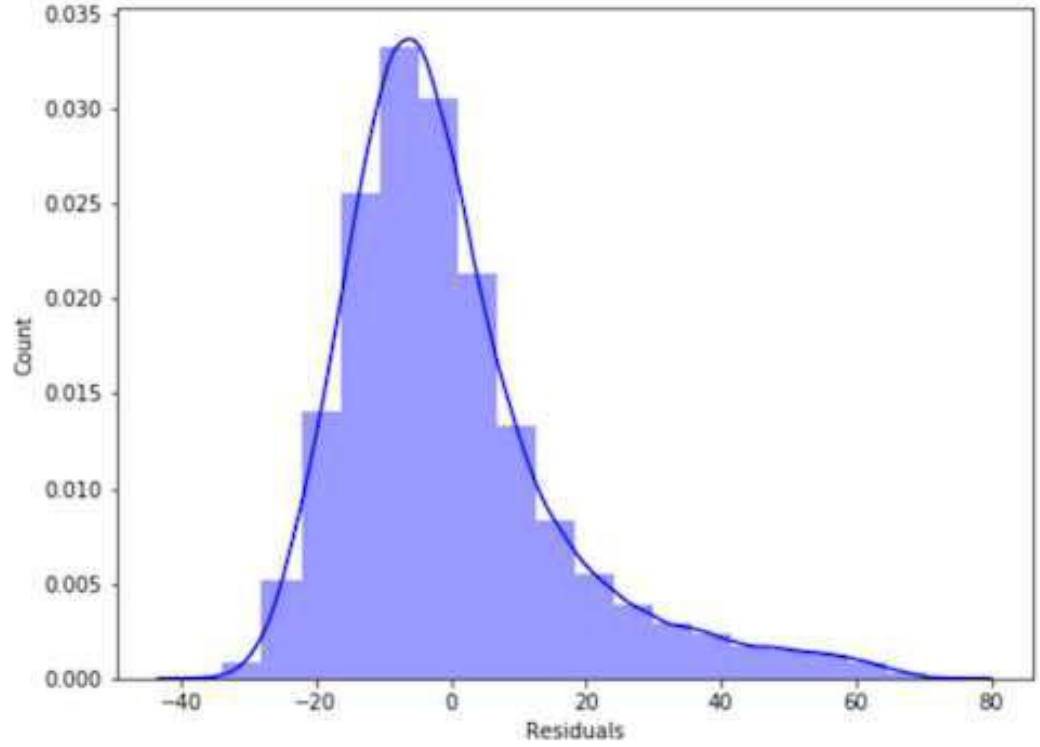
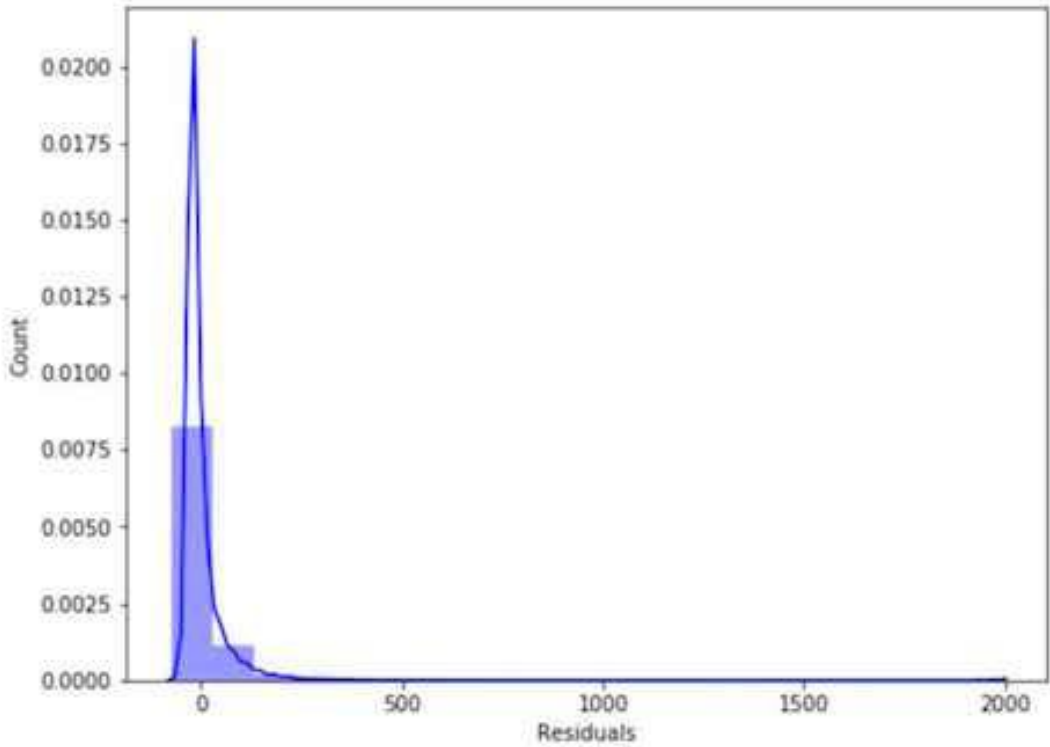
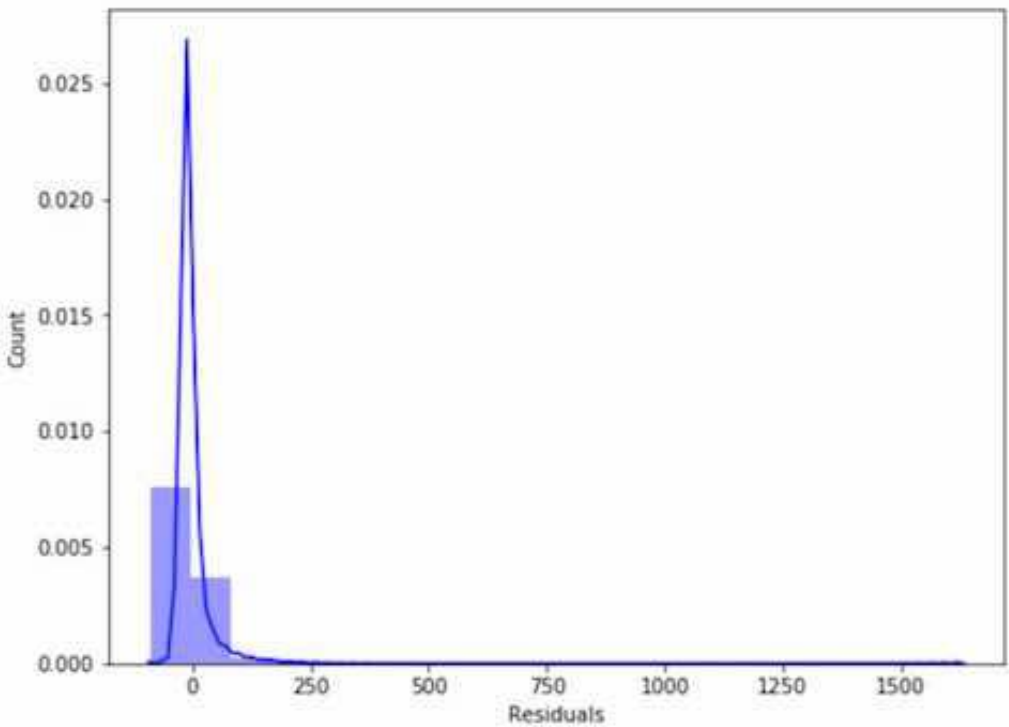
- loss = squared\_loss
- penalty = elasticnet
- alpha = 0,000001, l1\_ratio=0,05

- loss = squared\_loss
- penalty = l1
- alpha = 0,000001

- loss = squared\_loss
- penalty = elasticnet
- alpha = 0,0001, l1\_ratio=0,9

- loss = squared\_loss
- penalty = l1
- alpha = 0,000001

Résidus





# Choix du modèle final

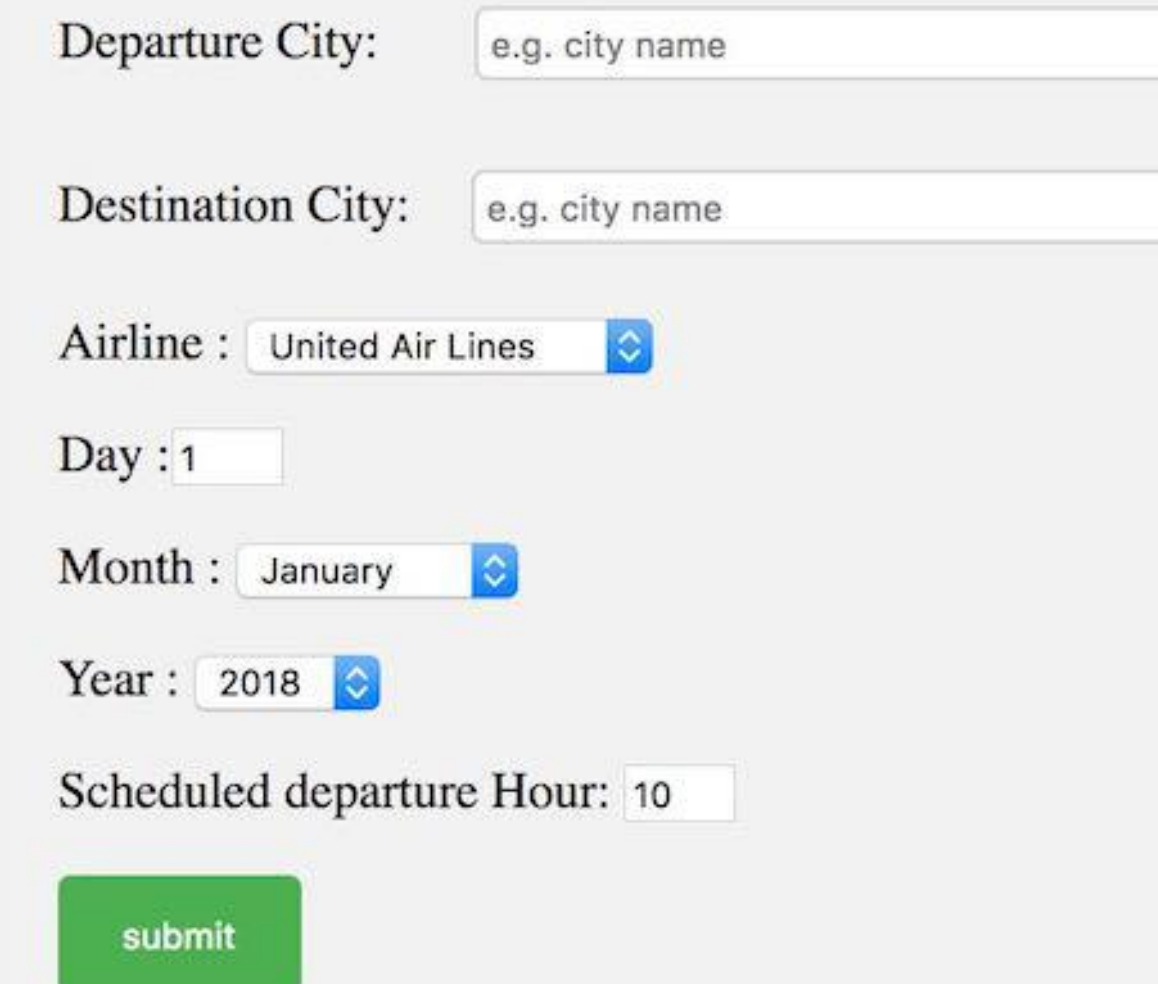
- ❑ L'équilibrage entre retards et avances et sans extrêmes donne la meilleur performance (MAE, MSE, Résidus)

## Delay Prediction

### Flight Information

- ❑ Modèle implémentée dans l'interface web :

<https://flightdelaypredicator.herokuapp.com/>



The screenshot shows a web form titled "Flight Information" for a "Delay Prediction" application. The form includes the following fields:

- Departure City:** A text input field with the placeholder "e.g. city name".
- Destination City:** A text input field with the placeholder "e.g. city name".
- Airline :** A dropdown menu currently showing "United Air Lines".
- Day :** A text input field containing the value "1".
- Month :** A dropdown menu currently showing "January".
- Year :** A dropdown menu currently showing "2018".
- Scheduled departure Hour:** A text input field containing the value "10".

At the bottom of the form is a green button labeled "submit".

# Conclusion

# Conclusion

- ❑ L'analyse exploratoire a permis une meilleure compréhension de nos données et d'identifier des variables pertinentes. Elle a aussi permis d'alléger la volumétrie de nos données.
- ❑ Nous avons amélioré les performances de notre modèle par des entraînements et la recherche des paramètres optimums.
- ❑ Le modèle a été implémenté via une interface web qui permet de prédire le retard ou l'avance d'un vol.
- ❑ Mais les retards exceptionnels restent difficile à prédire car souvent liés à des conditions météo ou des mesures exceptionnelles que l'on ne pourra jamais connaître à l'avance.

# Conclusion

- ❑ Comme nous l'avons vu dans nos graphes de comparaison entre valeurs mesurées et valeurs réelles, il y a encore un déficit dans le modèle et des pistes d'amélioration.
- ❑ L'ajout de données supplémentaires comme la météo serait intéressant à tester mais cela ne pourra améliorer que la prévision à court terme.
- ❑ D'autres algorithmes pourraient aussi être testés comme :
  - AdaBoostRegressor
  - Times series qui ajoutent aux données une dépendance à la dimension temps
  - Ou des algorithmes de type Classifieur (Logistic Regression, ...) pour dire si le vol aura ou pas du retard.