Anticipation des retards de vol d’avions

Adil DHISSA

Contexte et objectif

Base de données	API	Utilité
• https://www.transtats.bts.gov/ • Vols américains 2016	• Input: Données sur le vol • Output: Prédiction du retard du départ	• Anticiper les retards • Optimiser la logistique

Processus

Exploration

- Analyse
- Visualisation
- Nettoyage

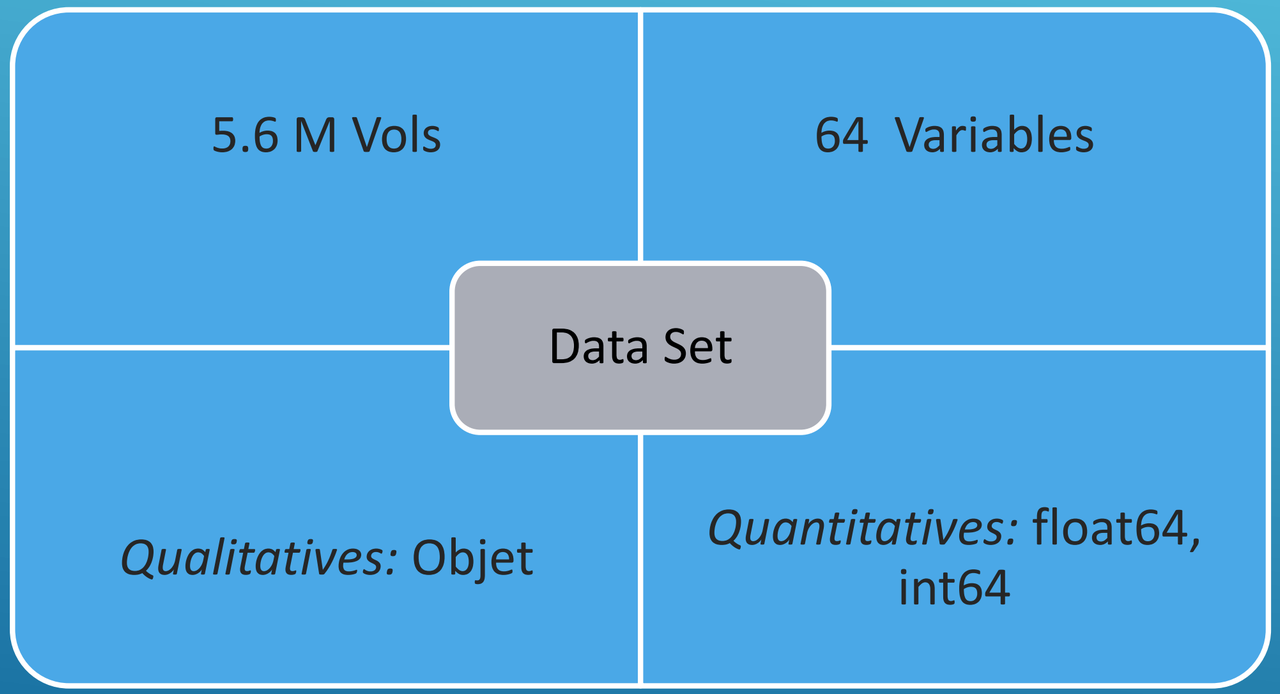
Modélisation

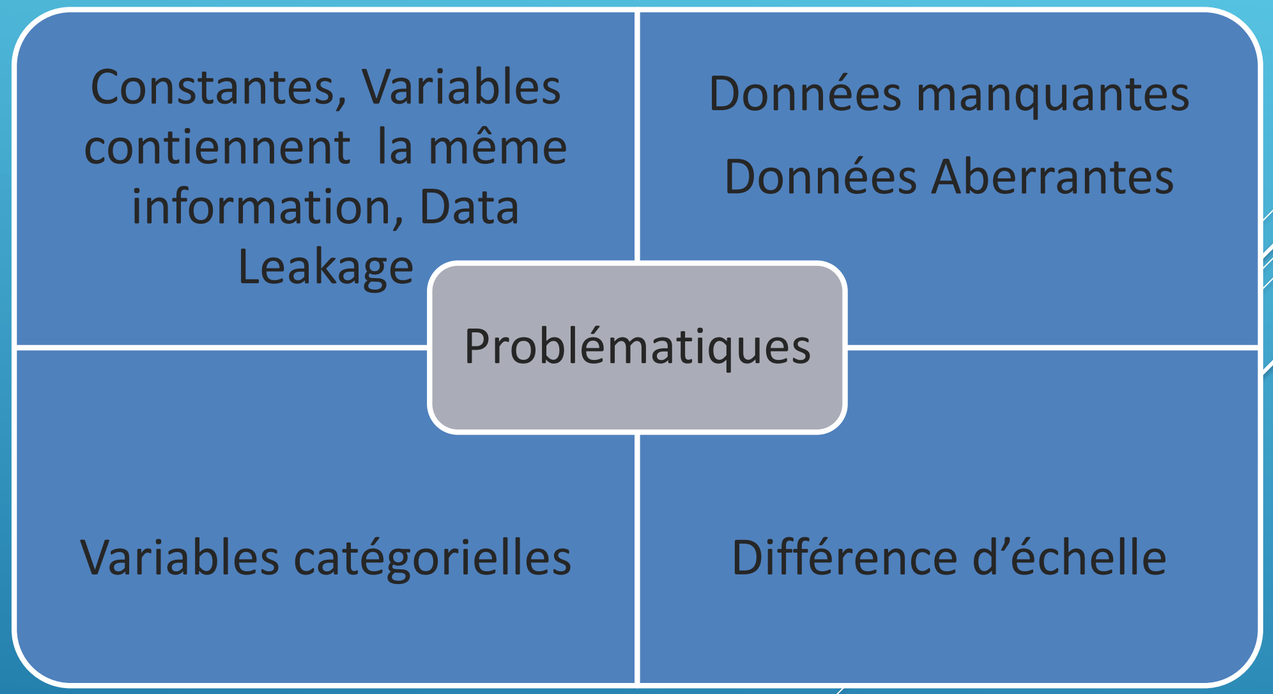
- Test des modèles supervisés
- Evaluation des modèles
- Choix du modèle final

API

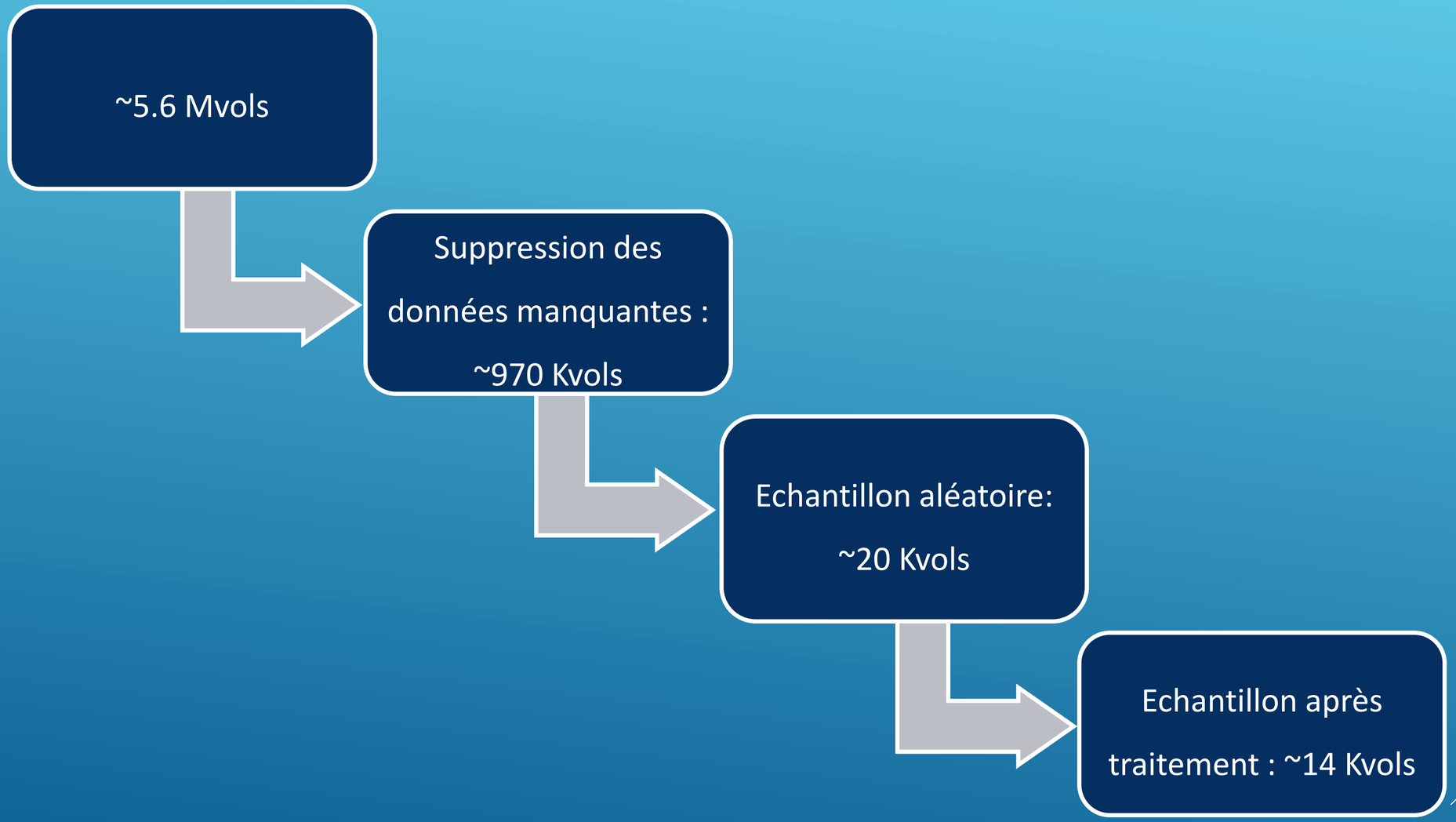
- Création d’une API de prédictions des retards de vols
- Déploiement sur pythonanywhere

Base de données



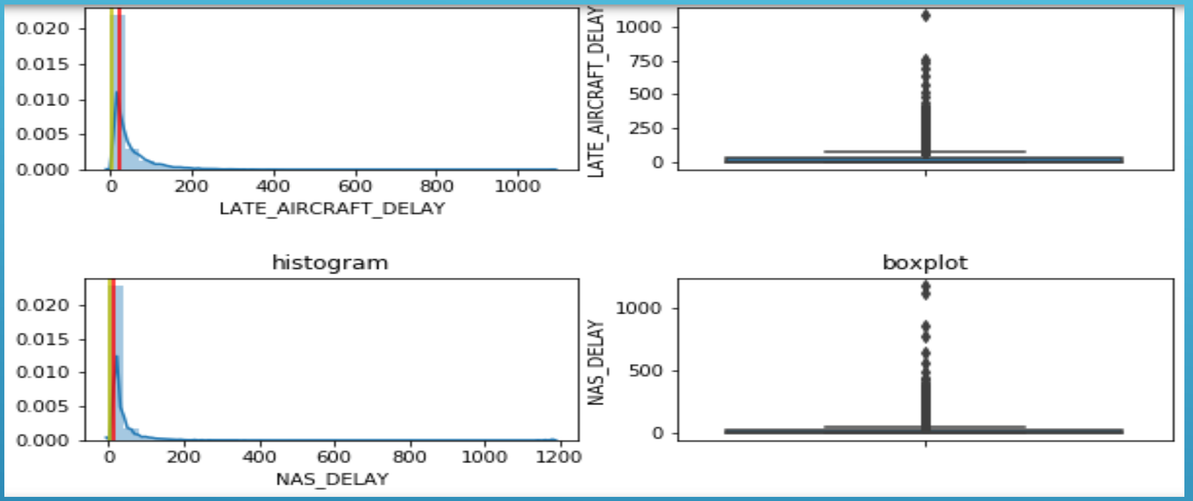


Échantillon de travail

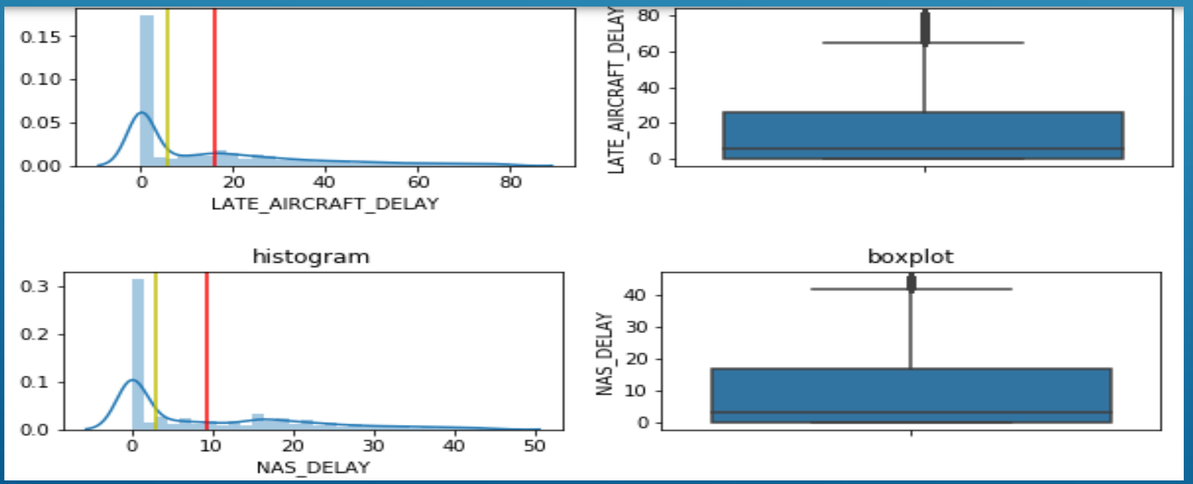


Traitement des variables numériques

Avant traitement des Outliers



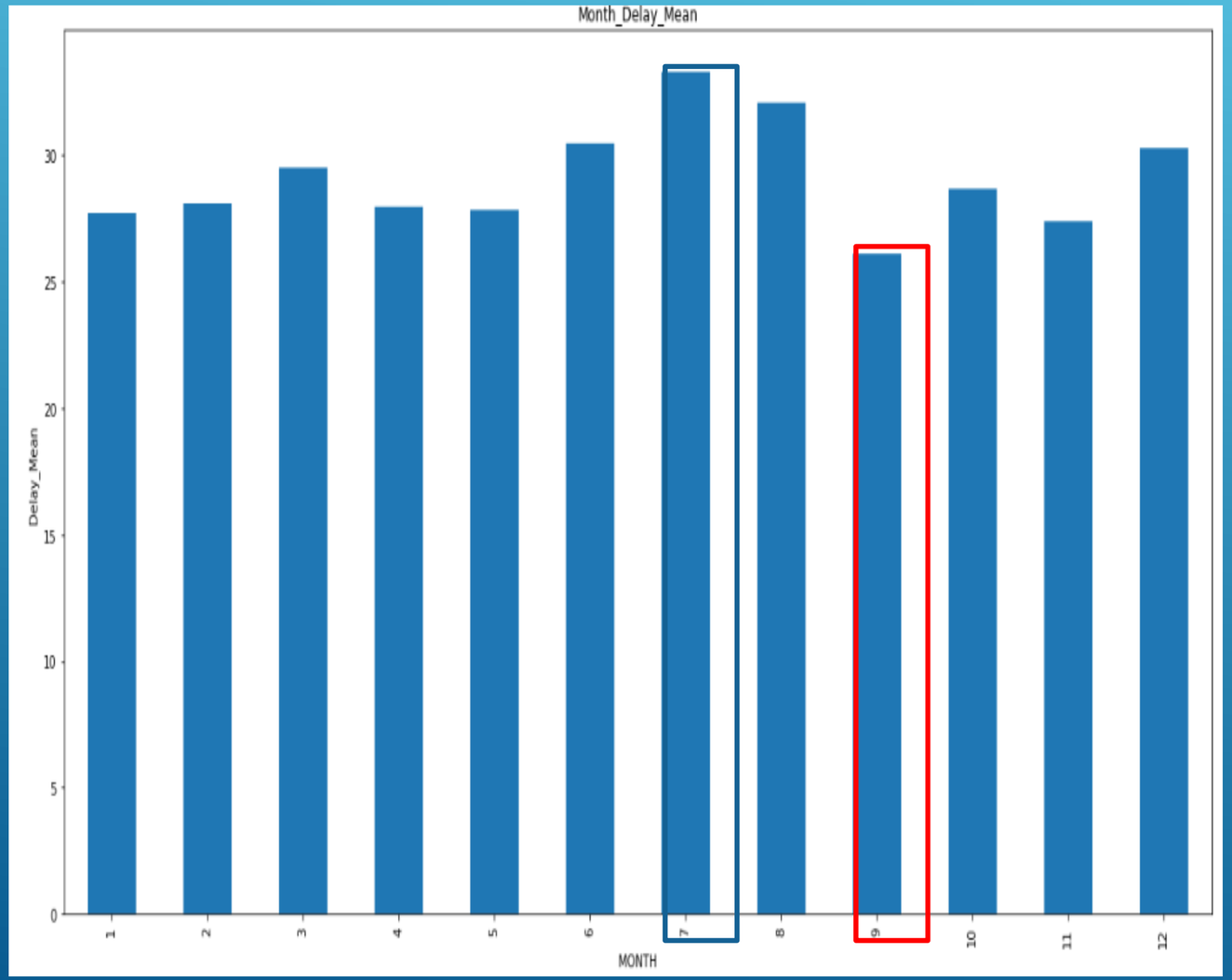
Après traitement des Outliers



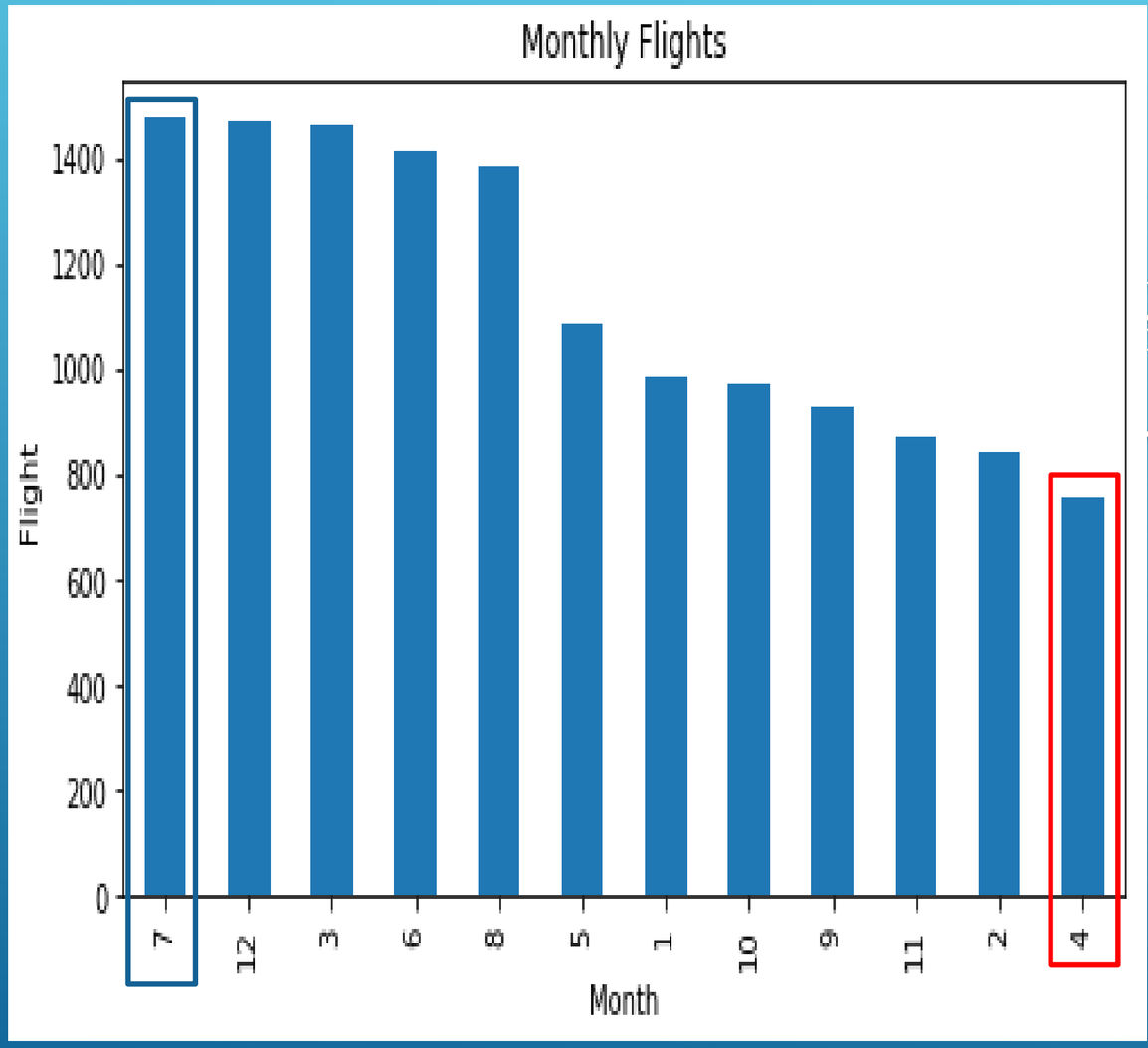
- Identification des outliers : Utilisation de L'écart inter-quartile
- L'écart inter-quartile est la différence entre le 3e quartile et le 1e quartile : IQ = Q3-Q1
- Outliers ~ Valeurs > Q3+ 1,5IQ et Valeurs < Q1- 1,5IQ
- Suppression des outliers

EXPLORATION

Moyenne des retards par mois

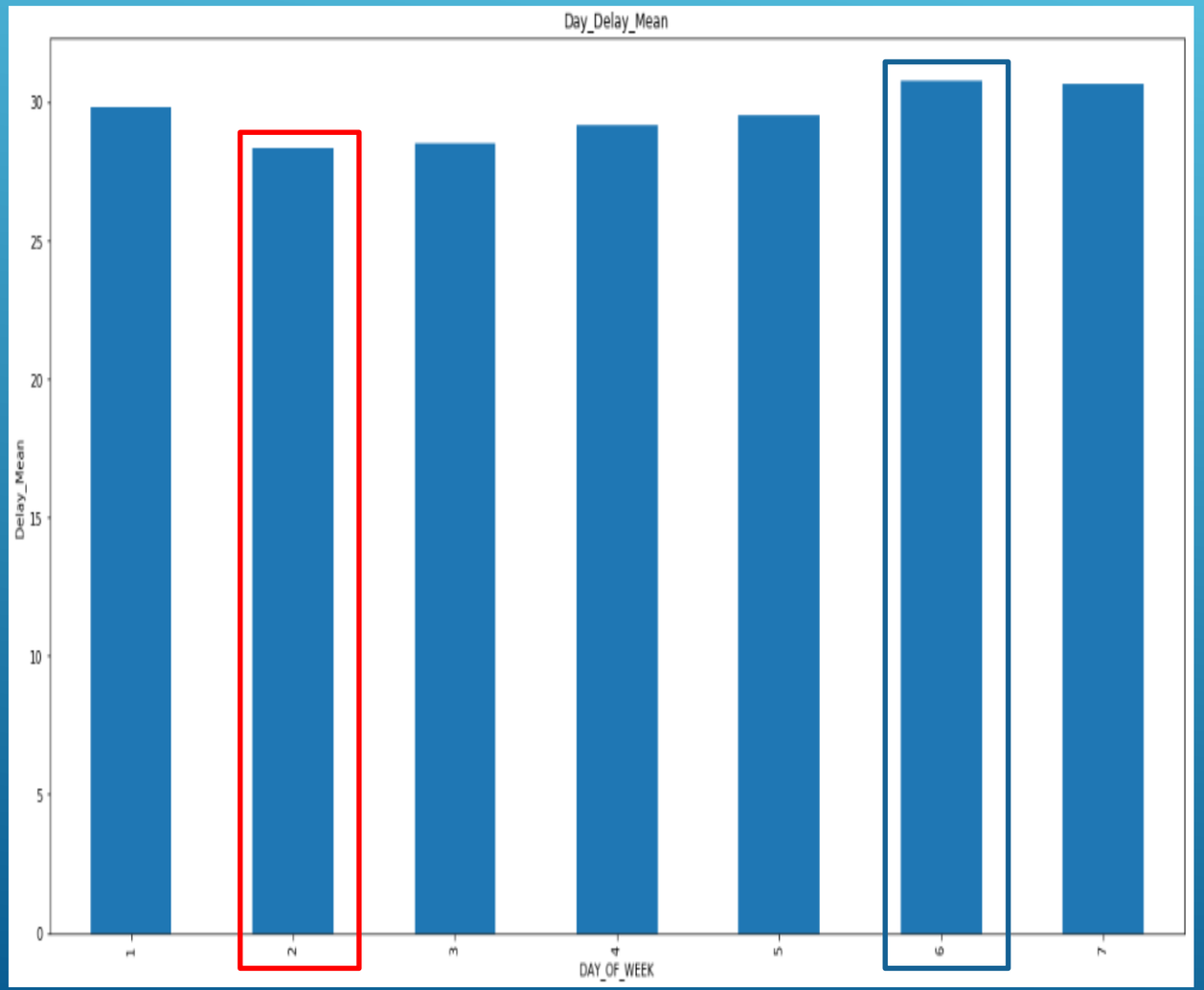


Nombre de vols par mois

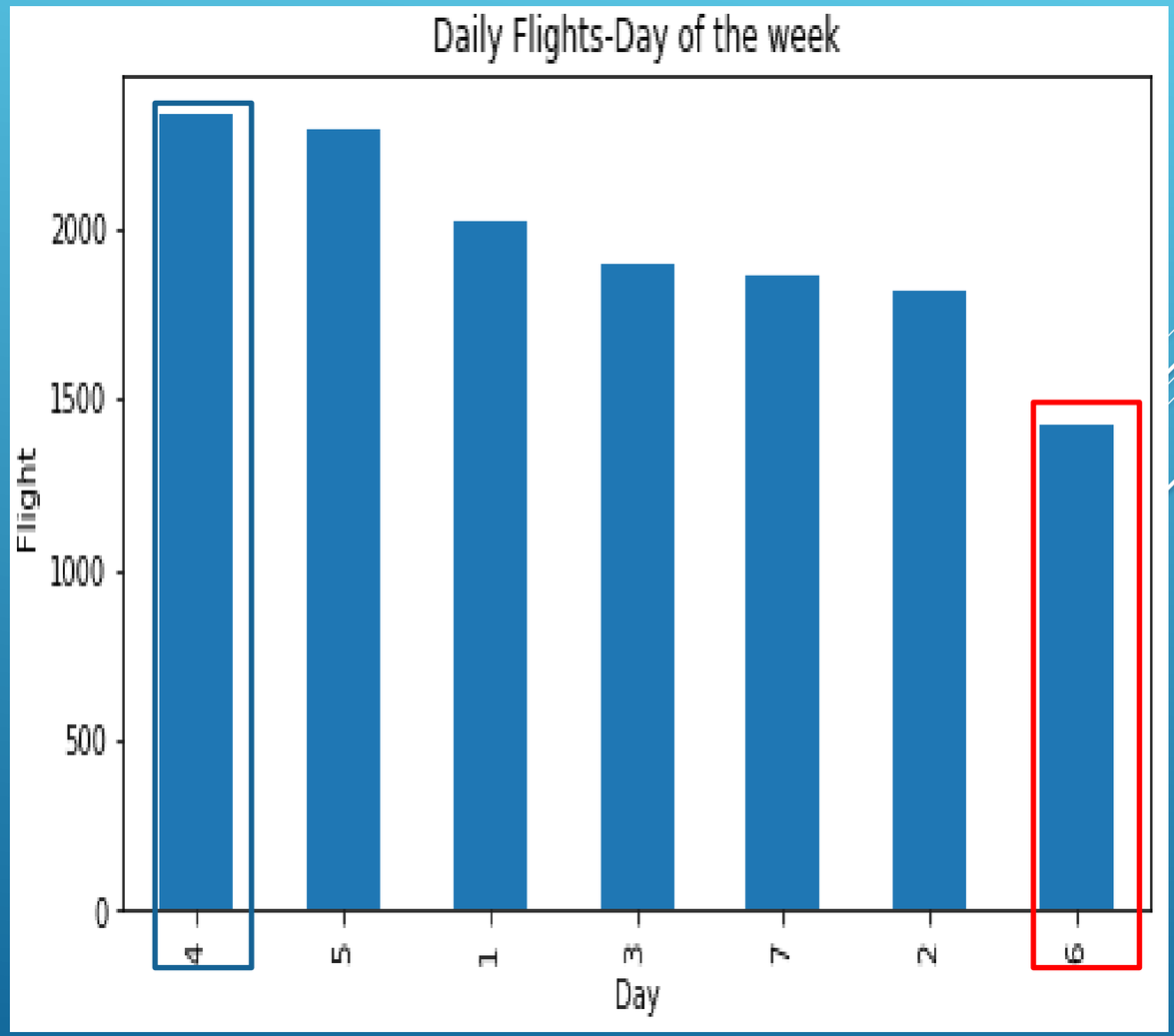


EXPLORATION

- Moyenne des retards par jour de la semaine

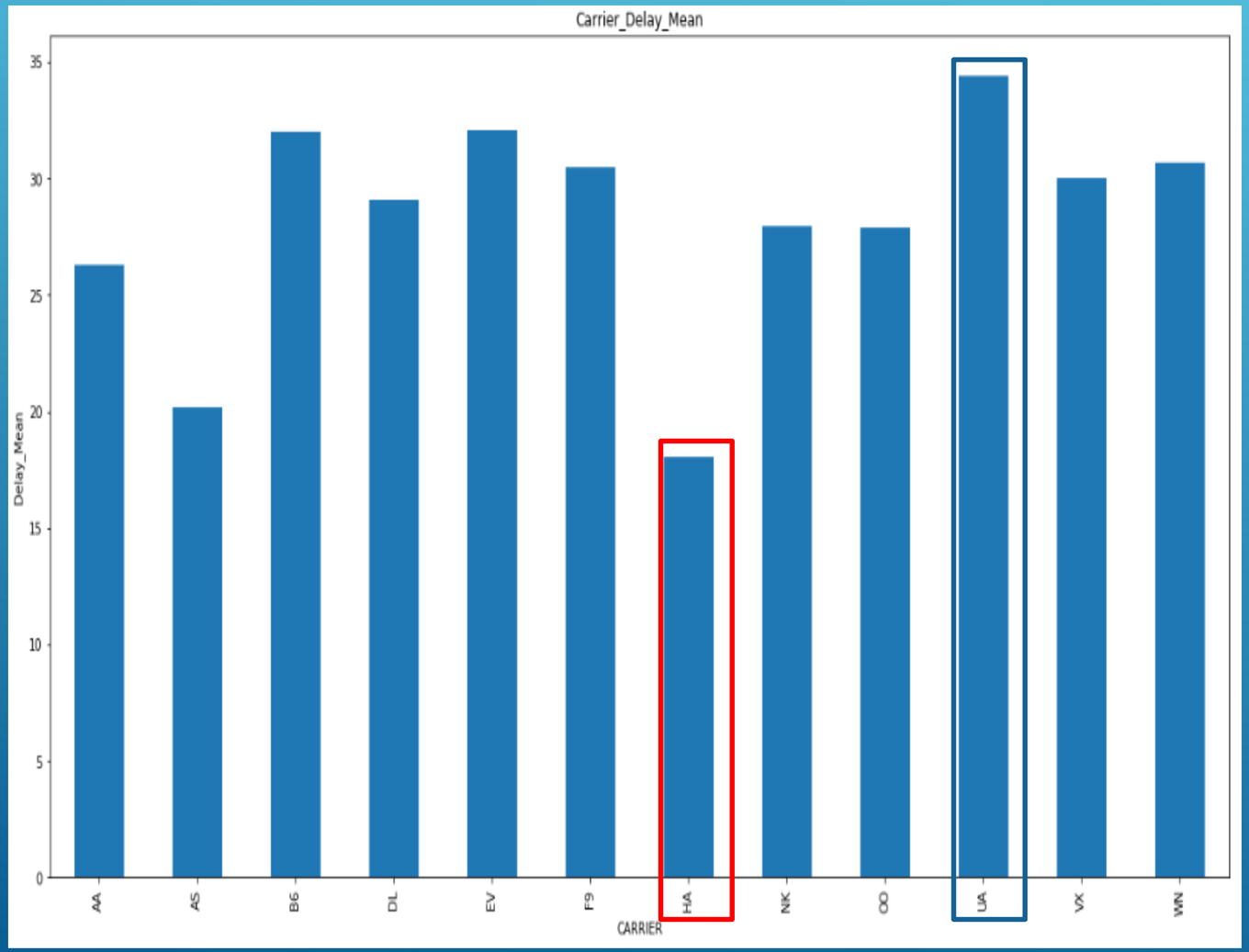


- Cumul des vols par jour de la semaine

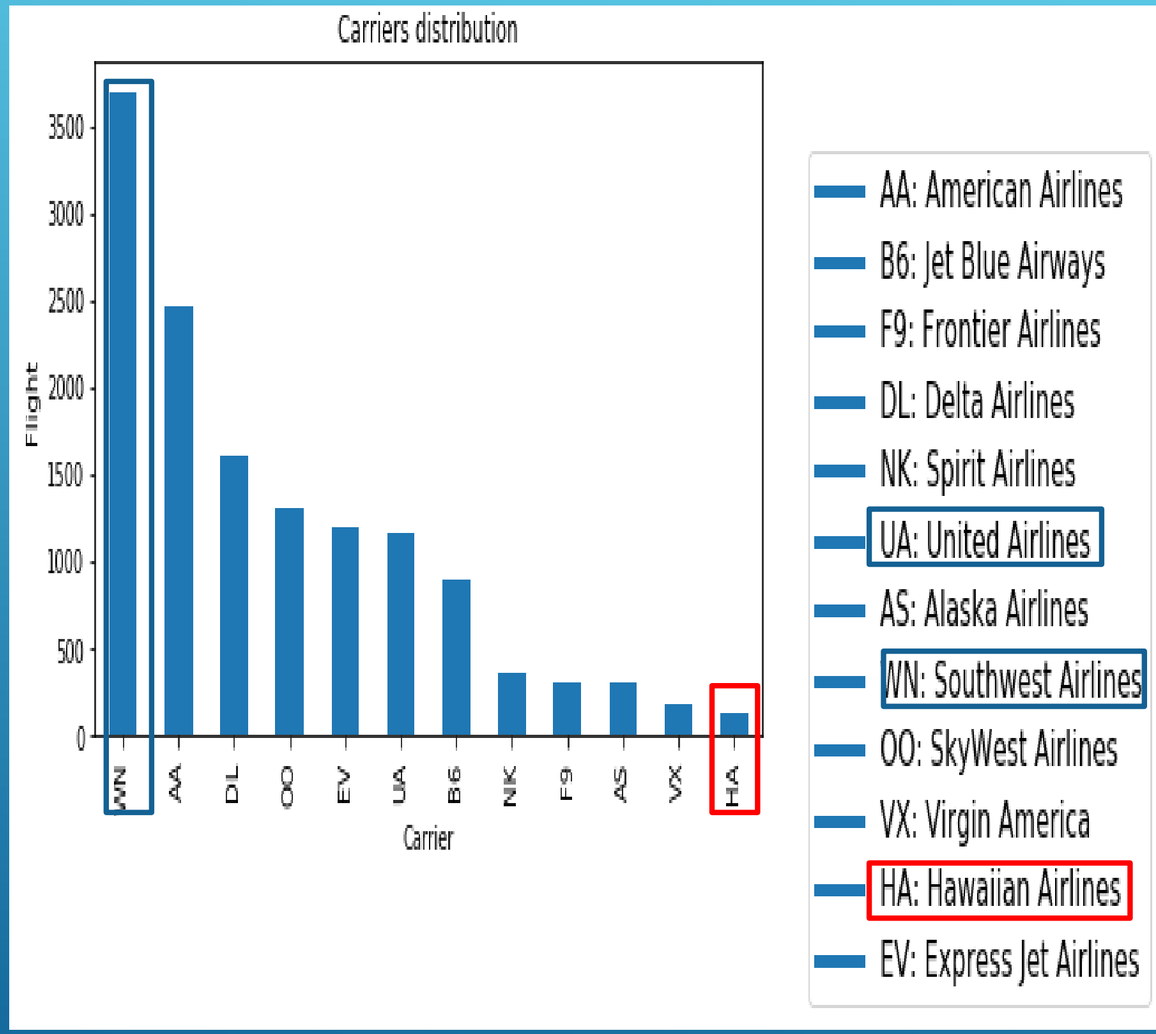


EXPLORATION

Moyenne des retards par compagnie

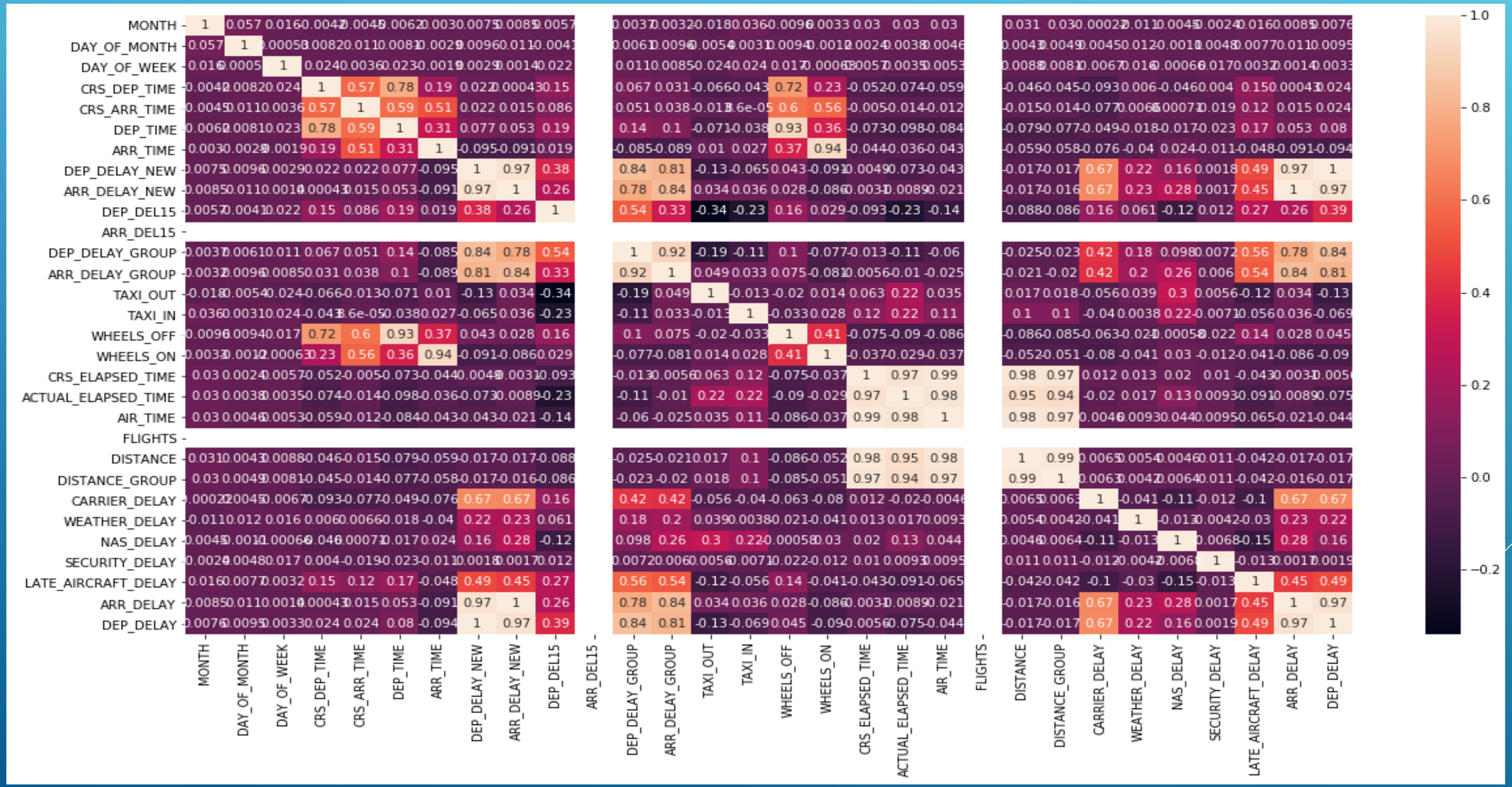


Nombre de vols par compagnie



EXPLORATION

La carte des corrélations



Sélection des variables

Critères de sélection des variables

Même Information	Leakage	Constance	Corrélation
▪ FL_DATE et YEAR, QUARTER, MONTH, DAY OF MONTH… ▪ UNIQUE_CARRIER et CARRIER ▪ Les variables DEST et ORIGIN, ORIGIN_CITY_NAME, DEST_CITY_NAME	▪ TAXI_OUT, TAXI_IN ▪ WHEELS_OFF, WHEELS_ON ▪ DEP_TIME, ARR_TIME, AIR_TIME…	▪ CANCELLED, DIVERTED, FLIGHTS…	▪ DISTANCE, DISTANCE_GROUP ▪ CRS_ELAPSED_TIME…

Variables DELAY (Cas particulier)

Inputs-Output API

Inputs:

- MONTH,
- DAY OF MONTH
- DAY OF WEEK
- CARRIER
- ORIGIN
- DESTINATION
- DEPARTURE TIME
- **Variables Delay (Cas particuliers)**

Output:

- DEP DELAY

➢ Traitement des variables catégorielles

CARRIER
AA
UA
WN
EV

CARRIER_AA	CARRIER_UA	CARRIER_WN	CARRIER_EV
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

- Les modèles ML sont basés sur des équations mathématiques
- Encoder les catégories en valeurs numériques
- Dummy Variable: Variable catégorique nominale à encoder par les « Dummy variables 0 ou 1)
- OneHotEncoding: MONTH, DAY OF MONTH, DAY OF WEEK, CARRIER, ORIGIN, DESTINATION, DEPARTURE TIME

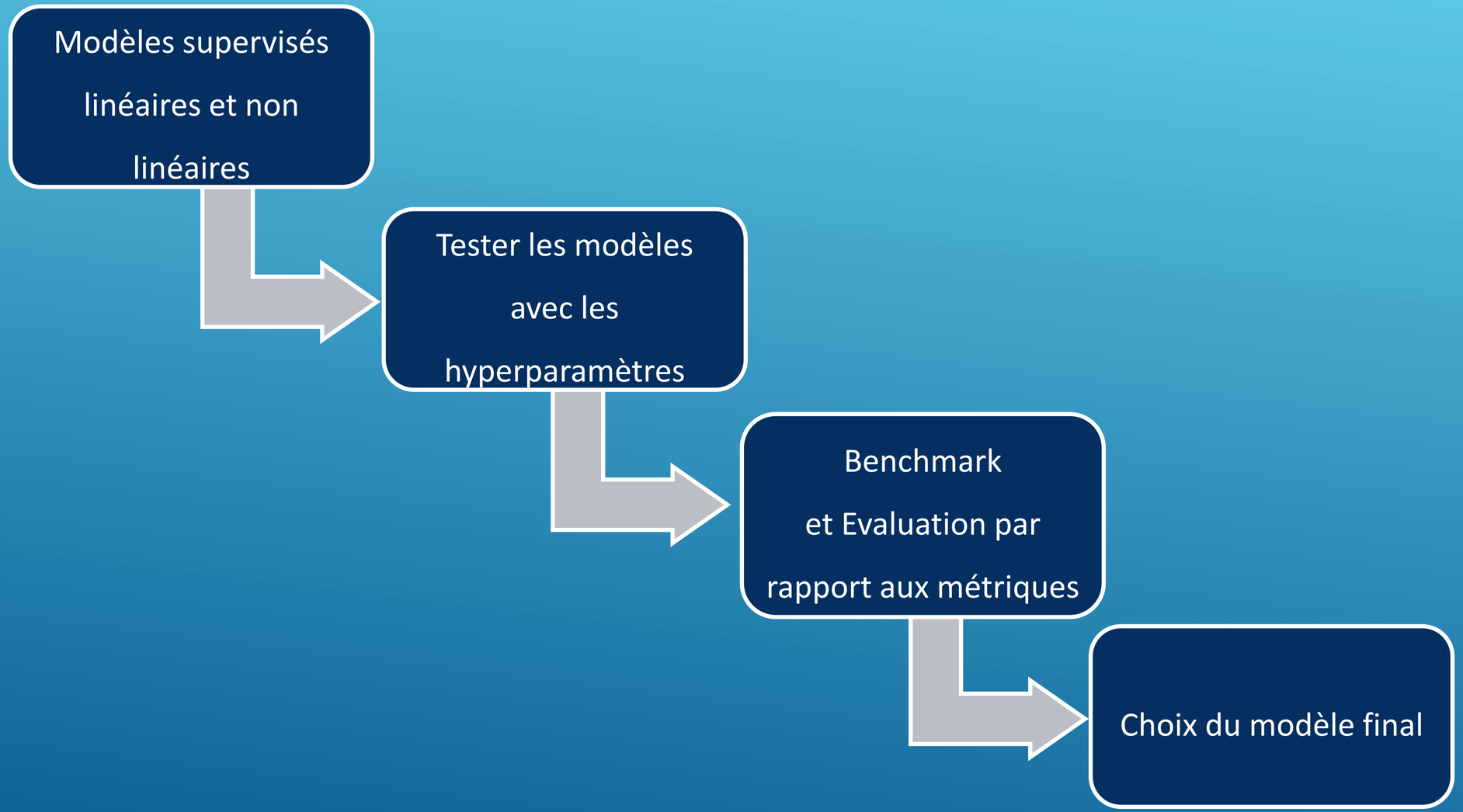
➢ Echantillon de modélisation prêt

	CARRIER_DELAY	WEATHER_DELAY	NAS_DELAY	SECURITY_DELAY	LATE_AIRCRAFT_DELAY	DEP_DELAY	CARRIER_AA	CARRIER_UA	CARRIER_WN
0	-0.720957	-0.065652	1.835812	-0.026497	-0.773717	-3	1	0	0
1	2.741342	-0.065652	2.683872	-0.026497	-0.773717	39	0	1	0
2	1.054581	-0.065652	-0.623562	-0.026497	-0.773717	20	0	0	1
3	-0.720957	-0.065652	0.394110	-0.026497	0.056031	17	0	0	0
4	0.877027	-0.065652	-0.793174	-0.026497	-0.773717	20	0	1	0

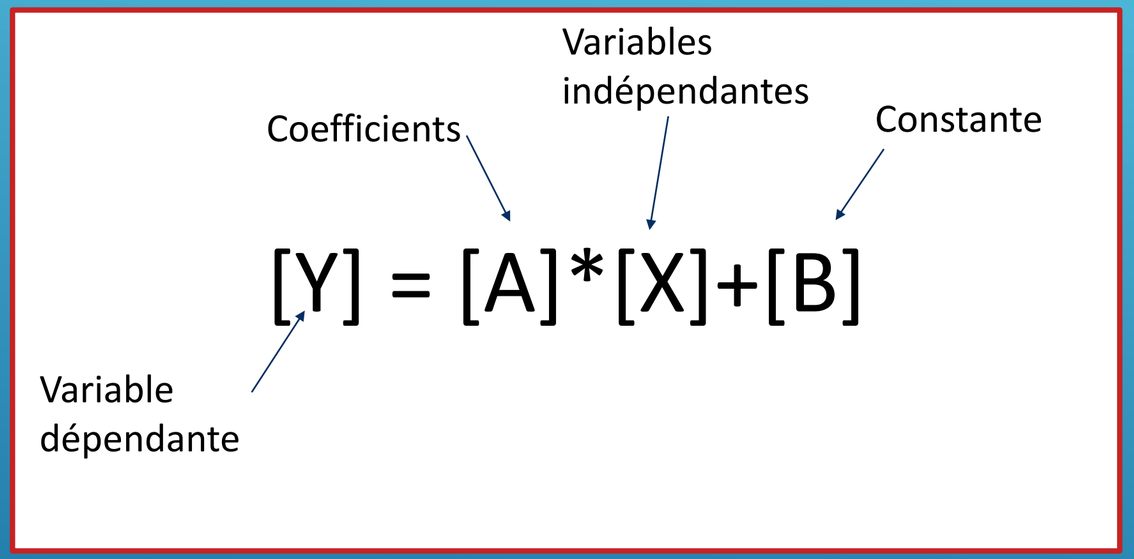
5 rows × 633 columns

- 13645 vols
- 633 variables
- Les variables « DELAY » standardisées
- « DEP_DELAY » variable dépendante non standardisée
- Les variables catégorielles non standardisées (0,1)

➢ Modélisation

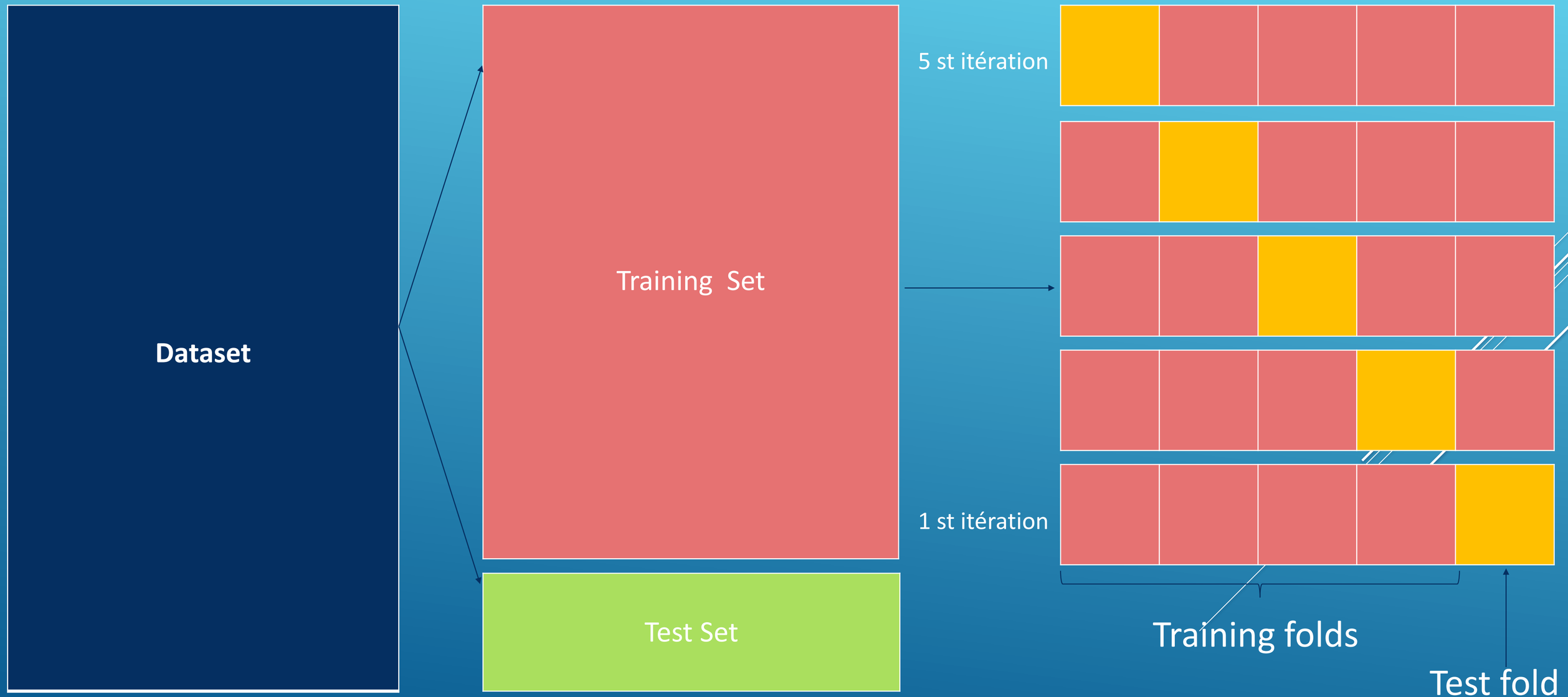


Régression linéaire multiple

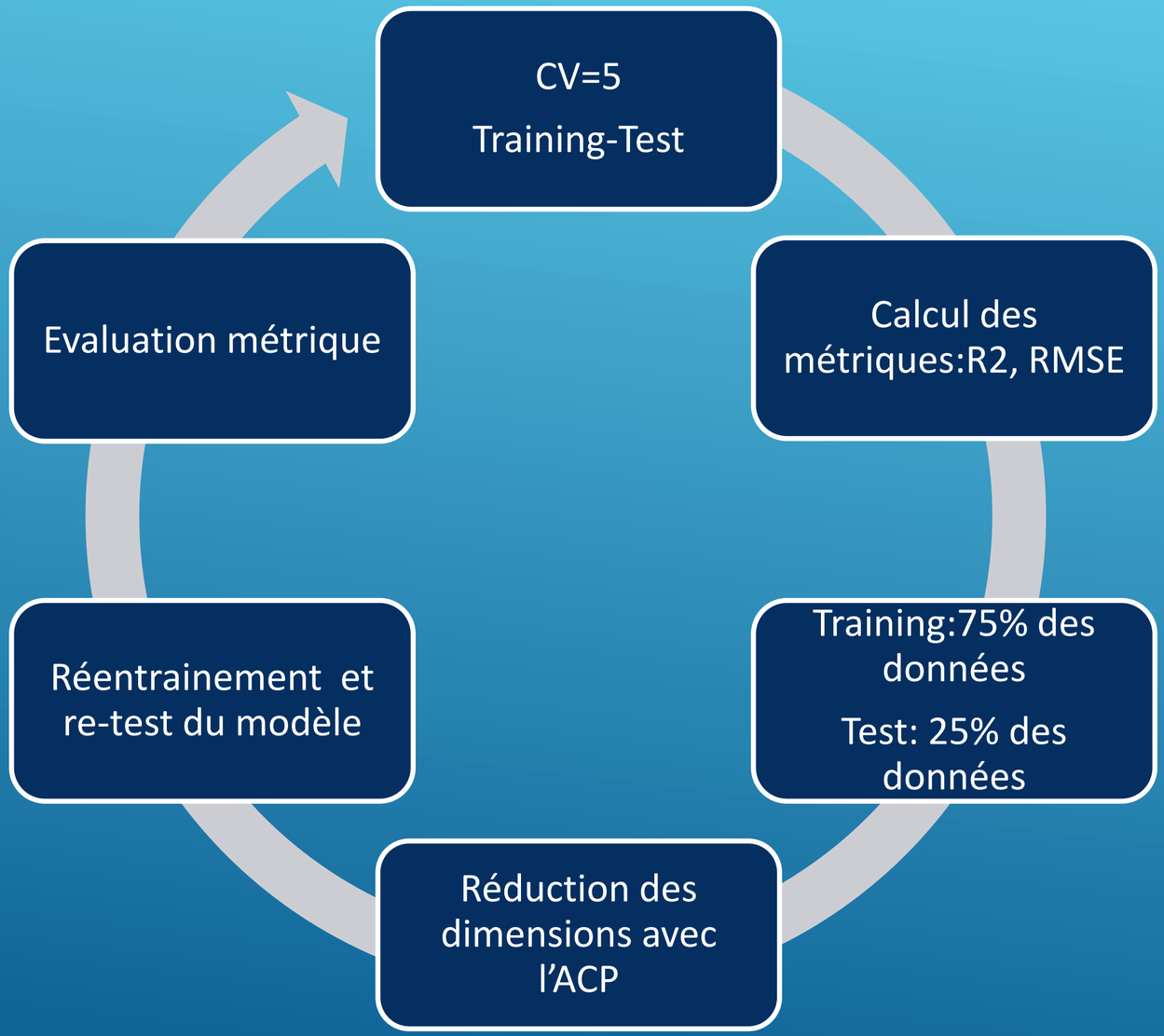


- Méthode d'apprentissage supervisé
- La variable à prédire « DEP_DELAY » dépend linéairement des variables indépendantes (Date et heure du départ, Aéroports…)

Validation Croisée



➢ Application RLM+ACP



➢ Résultats RLM + ACP

Métriques	Résultats
R^2	-0,075
RMSE	24,83

- RMSE et R^2 mesurent la performance du modèle
- Si RMSE=25 minutes , on peut s'attendre à une valeur de y_pred décalée de 25 min en moyenne
- R^2: 0 <= R^2 <=1
- R^2=1 veut dire y_pred = y_test: Le modèle est parfait
- Cas R^2<=0 : Les valeurs y_pred sont très loin des y_test
- ACP Reduction des dimensions => Perte d'information

Régularisation

Problématique:

-Grand nombre de variables

-Variables corrélées

-Capture du bruit produit par les données

-Coefficients instables

-Variance élevée - Overfitting

-Petit nombre de variables

-Biais élevé - Underfitting

Solution:

Régularisation:

-Contrôler la complexité du modèle: Ajouter un terme de régularisation à RSS

-Ridge:

Min (SUM(Y-Yreg)² + $\alpha * slope^2$)

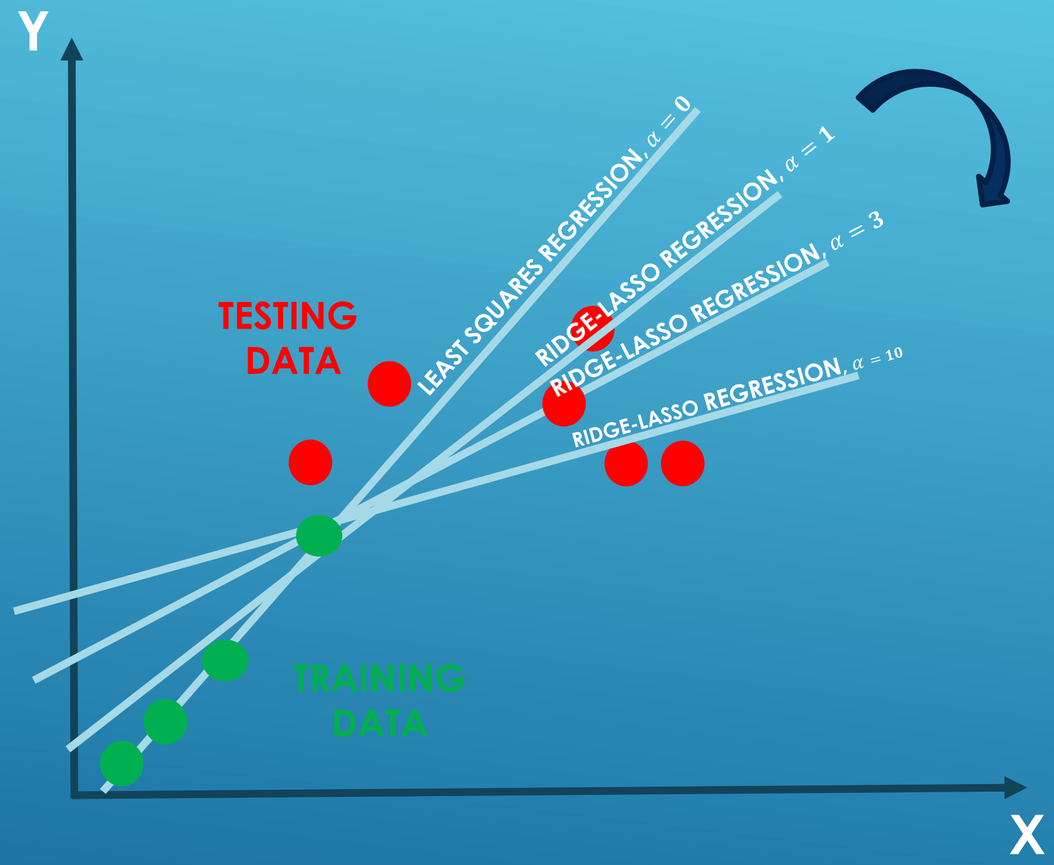
Limiter l'overfitting et réduire les poids près de 0 (Mais pas exactement à 0) => Pas de suppression des variables

-Lasso:

Min (SUM(Y-Yreg)² + $\alpha * |slope|$)

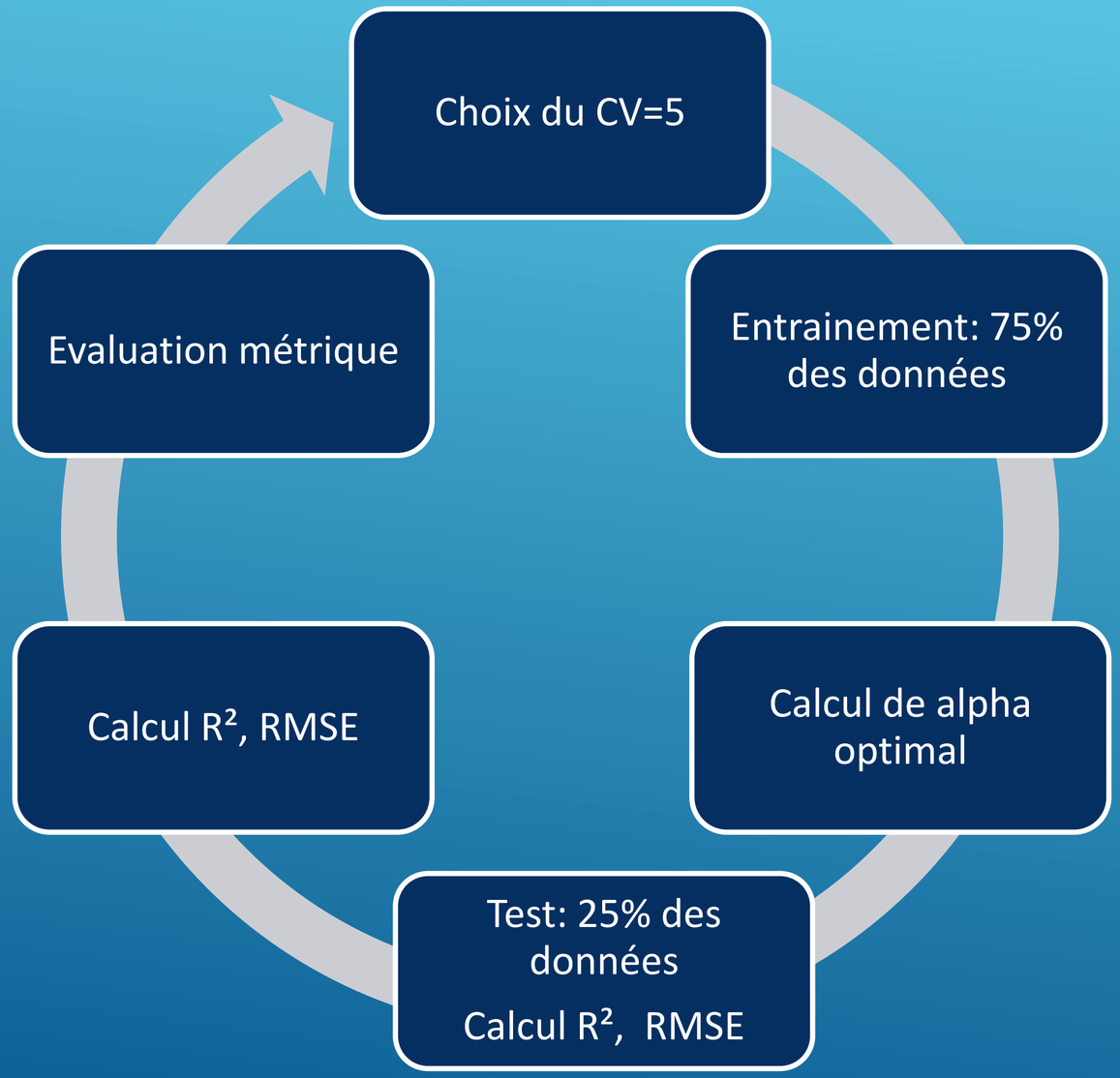
Limiter l'overfitting et possibilité de réduire les poids à 0=>Suppression des variables

Régression Ridge-Lasso



- Overfitting: Le modèle trop spécialisé sur les données du training set et qui se généralisera mal sur les données test set
- En changeant la pente, Ridge-Lasso essayent d'augmenter le biais pour améliorer la variance => Améliorer la capacité de généralisation du modèle
- Si α est très grand la pente devient très petite et le modèle devient moins sensible aux changements des variables indépendantes=> Underfitfing

Application Ridge-Lasso



Résultats Ridge-Lasso

Régression Ridge

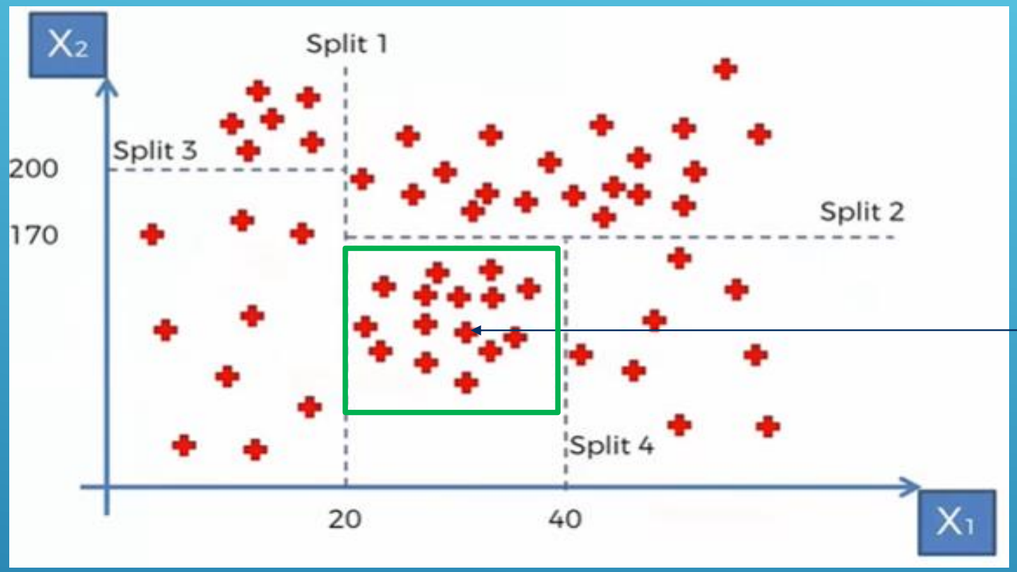
Métriques	Résultats
R^2	0,062
RMSE	23,69

Régression Lasso

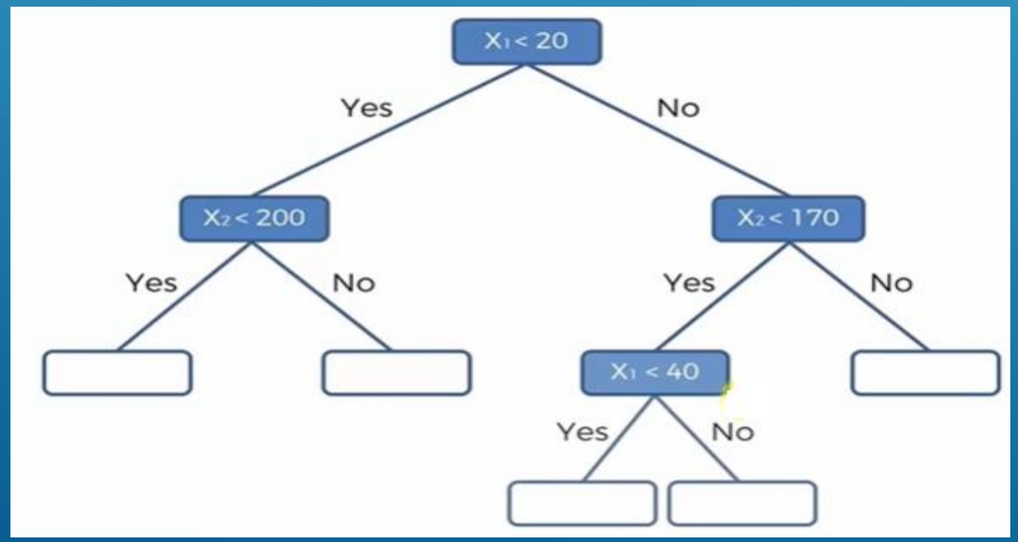
Métriques	Résultats
R^2	0,062
RMSE	23,69

- Ridge et Lasso ont les mêmes performances
- $R^2>0$, Mais reste très bas
- RMSE élevées
- Mauvaises performances des deux modèles

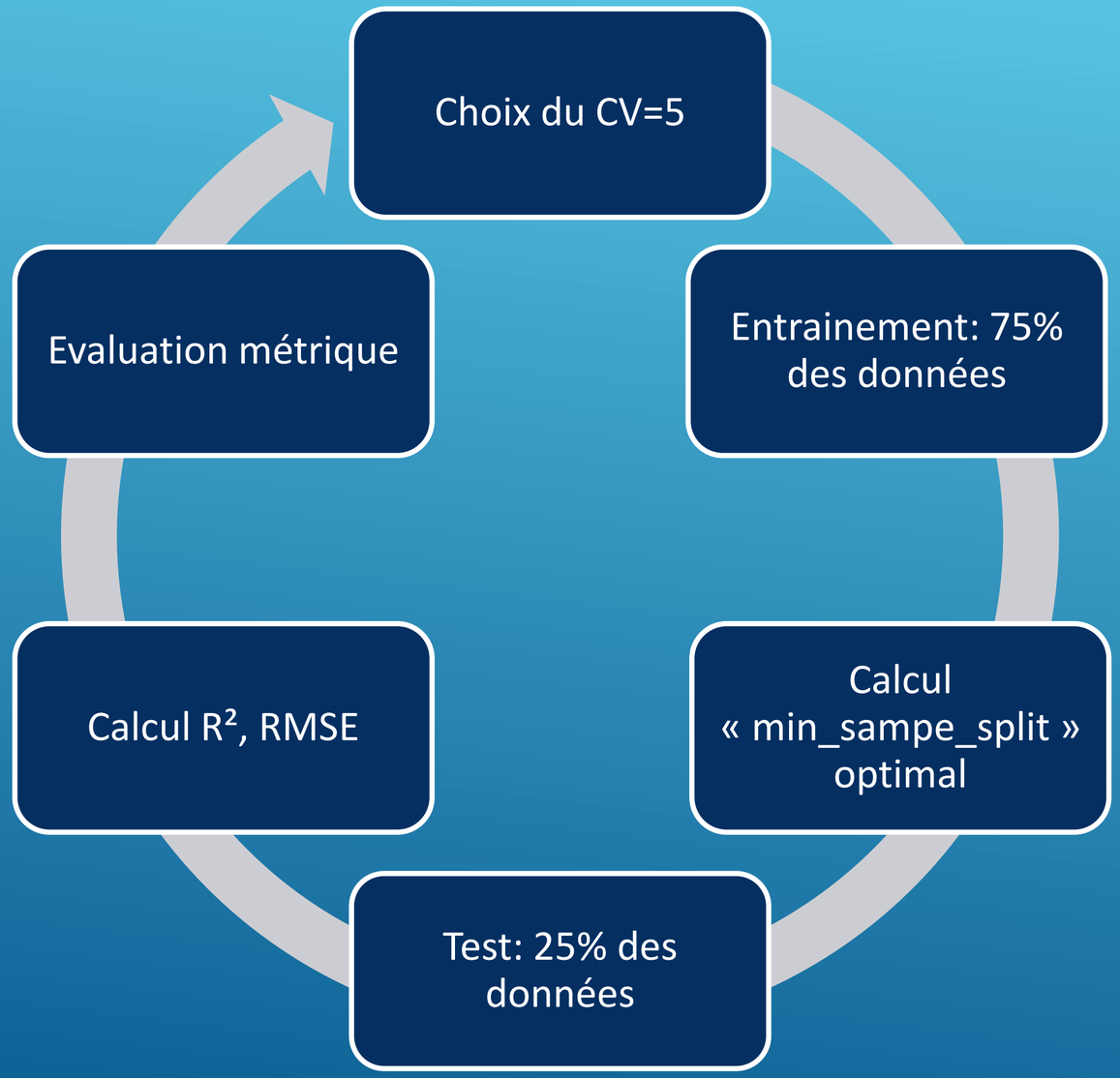
Decision-Tree



- Méthode d'apprentissage supervisé
- Diviser le jeu de données en plusieurs splits
- La prédiction de Y de l'observation (X1=30, X2=100) est la moyenne des toutes les valeurs appartenant au split contenant (X1,X2)



➢ Application Decision-Tree



➢ Résultats Decision-Tree

Métriques	Résultats
R^2	-1,23
RMSE	28,96

- $R^2<0$, les valeurs DEP_DELAY_pred sont très loin des DEP_DELAY
- RMSE élevée
- Decision-Tree est moins performant que Ridge et Lasso

➢ Booster les performances des modèles

Problématique:

-Mauvaises performances de l'ensemble des modèles

-Mauvais modèle => mauvaise prédiction

-Pas de valeur ajoutée pour l'utilisateur final

Solution:

-Utilisation des variables DELAY

-Réentraîner et retester les modèles avec les variables DELAY

-Evaluation métrique

API:

-Inputs initialement sélectionnées

-KNN intermédiaire : Estimation des valeurs DELAY en utilisant les mêmes inputs de l'API

-Intégrer les valeurs DELAY aux inputs de l'API

-Prédire DEP_DELAY

Choix Final du modèle

Résultats des modèles sans variables DELAY

Métriques	Ridge	Lasso	Decision-Tree
R^2	0,062	0,062	-1,23
RMSE	23,69	23,69	28,96

Résultats des modèles avec les variables DELAY

Métriques	Ridge	Lasso	Decision-Tree
R^2	0,85	0,85	0,82
RMSE	9,50	9,47	10,43

- Ridge et Lasso ont les mêmes performances
- Ridge et Lasso sont meilleurs par rapport aux autres modèles
- Ridge est plus rapide lors de l'exécution et garde toutes les variables

Sélection des variables DELAY

Impact des variables Delay

Métriques	Sans DELAY	CARRIER	CARRIER+ WEATHER	CARRIER+WEATHER+ NAS	CARRIER+WEATHER+ NAS+SECURITY	CARRIER+WEATHER+NAS+SECURITY+ LATE_AIRCRAFT
R^2	0,062	0,14	0,14	0,32	0,31	0,83
RMSE	23,69	23,03	22,73	20,21	19,82	10,06

Variables Delay Utiles

Métriques	CARRIER+NAS+ LATE_AIRCRAFT
R^2	0,84
RMSE	9,88

- Nette amélioration des métriques en intégrant les variables DELAY
- R^2 et RMSE ne s'améliorent pas en intégrant WEATHER_DELAY et SECURITY_DELAY
- Les variables DELAY utiles pour le modèle: CARRIER_DELAY, NAS_DELAY et LATE_AIRCRAFT_DELAY

Création et déploiement API

Environnement de développement

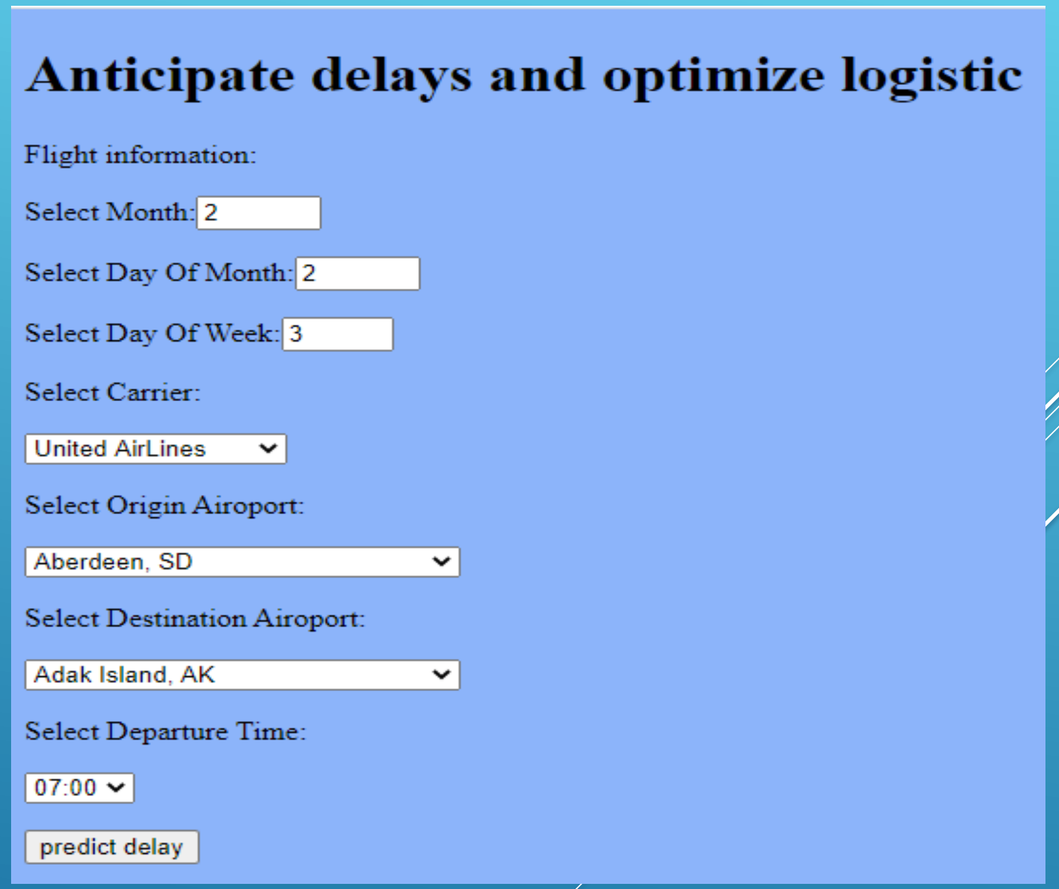
- Modèle RIDGE et une couche KNN pour estimer les valeurs DELAY
- BDD: Origin, Destination
- Virtual Env python
- Framework Flask et HTML

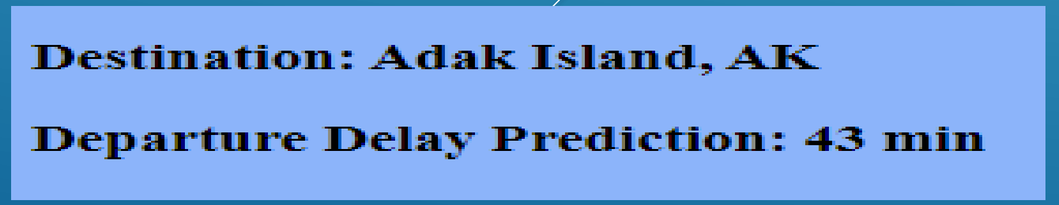
Serveur Web Local

- Pages HTML 1: Informations Vol
- Pages HTML 2: Prédiction du Retard

Environnement de production

- Pythonanywhere.com
- API en ligne: http://adildhissa.pythonanywhere.com/





Conclusion et perspectives

Conclusion

- La corrélation entre les variables et le Data leakage ont limité le choix des variables
- Mauvaise performance des modèles
- Nette amélioration des résultats après l'intégration des variables DELAY

Amélioration

- Elargissement de la base de données sur plusieurs années
- Plus d'information sur CARRIER, AIRCRAFT et NAS
- Classification des retards pour mieux définir le niveau et les moyens d'intervention des opérateurs