

# Projet Big data

## Prédiction du tarif des taxis à New York

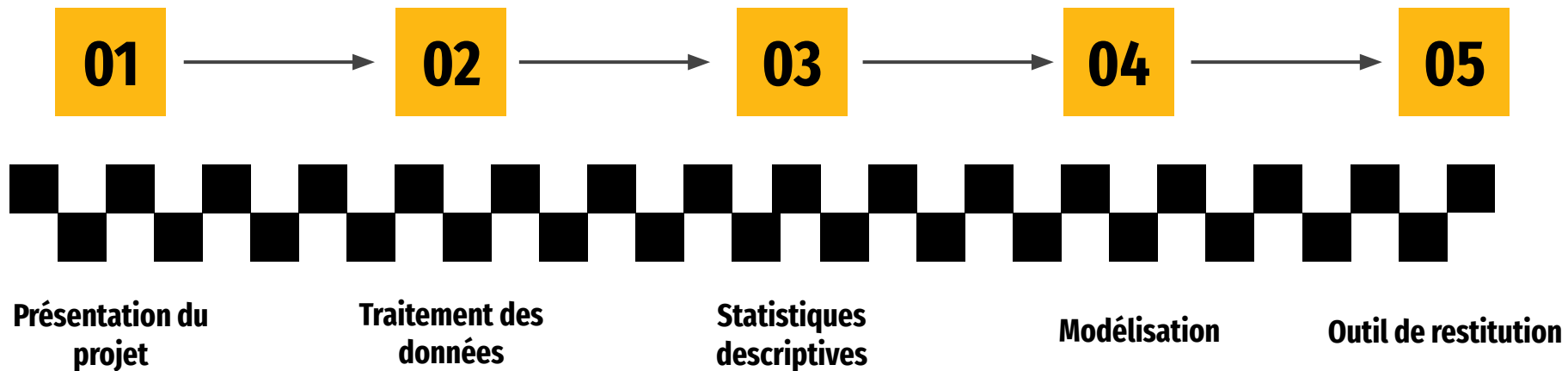


### Réalisé par :

ABDEL MOUTALEB Amine  
ADOKPE Komi  
BELHOTI Reda  
KA Racky

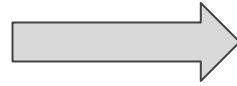
27/03/2023

# Sommaire

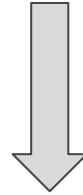
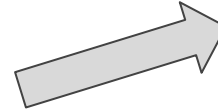


# Présentation du projet

## ❖ Contexte



DATASET



→ **Problématique** : prédire le tarif des taxis à New York.



# Hypothèses sur le tarif des courses en taxi

**Distance parcourue**



**Heure**



**Jour de la course**



**Lieu**



**Durée du trajet**



**Conditions  
météorologiques**



# Description du jeu de données

## Source

kaggle

Plateforme web de compétitions en data science.

## Fichiers

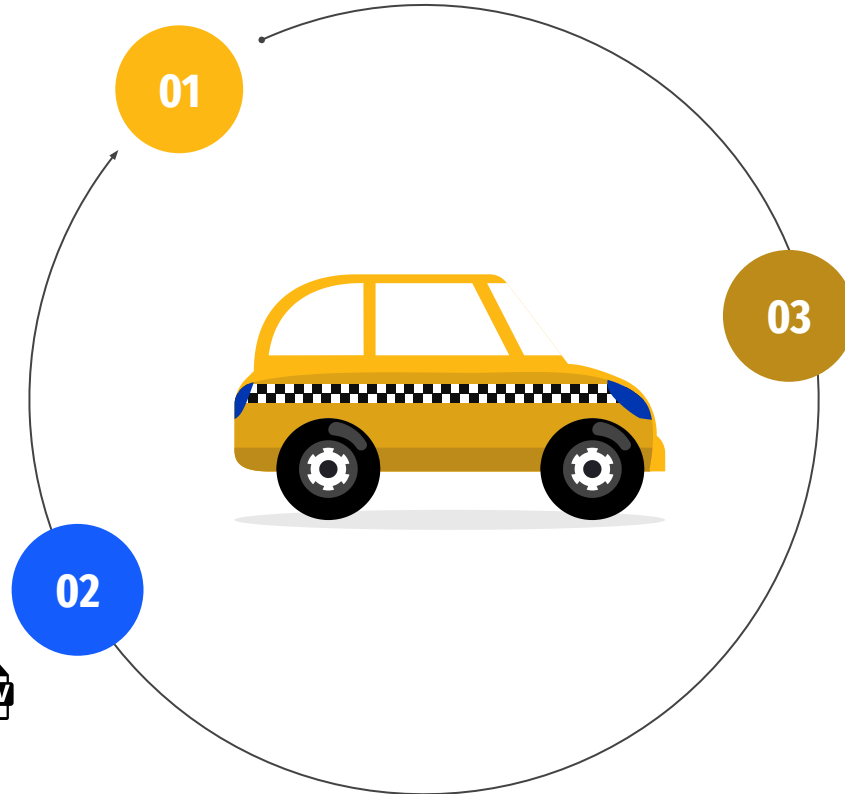
2 fichiers :

train.csv (jeu de données)

test.csv (données à prédire)

train.csv (55M de lignes)

test.csv (10K de lignes)



## Variables explicatives

- date de départ
- longitude de départ
- latitude de départ
- longitude d'arrivée
- latitude d'arrivée
- nombre de passagers

## Variable à expliquer

- tarif de la course



## Variable ID

- key



# Traitement de données

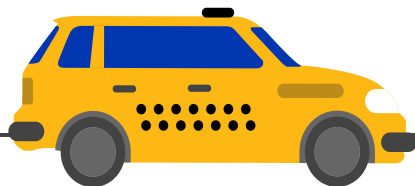
01

02

03

04

05



- **Nettoyage du jeu de données**
- **Création de variables**

# Traitement de données

## Tarif de la course

Trajets avec des tarifs incohérents (tarif < 2,5 \$US, le tarif forfaitaire de base ou tarif < 3,5 \$US pour des trajets en heures de pointe).

## Nombre d'observations

Avant traitement : 150000  
Après traitement : 122847

## Nombre de passagers

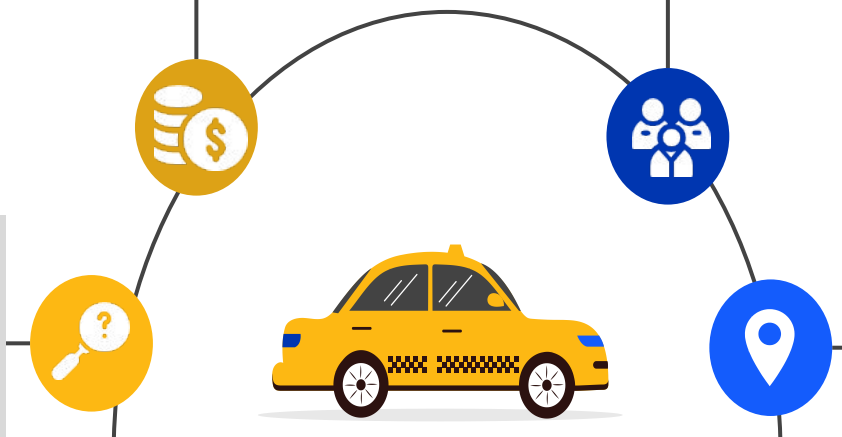
Trajets sans passager.

## Données manquantes

1 observation avec 2 valeurs manquantes.

## Longitudes et latitudes

Trajets qui sont réalisés en dehors de New York.  
Trajets dont les coordonnées de départ sont égales aux coordonnées d'arrivée.

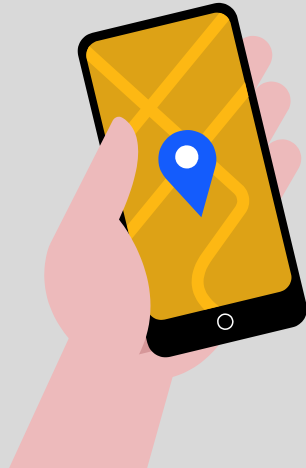


# Création de variables

01

## Distance de Manhattan

Création d'une variable **"distance"** à partir des coordonnées de longitude et de latitude.



02

## Variables de temps

Création des variables de temps à partir de la **date de départ de la course**.

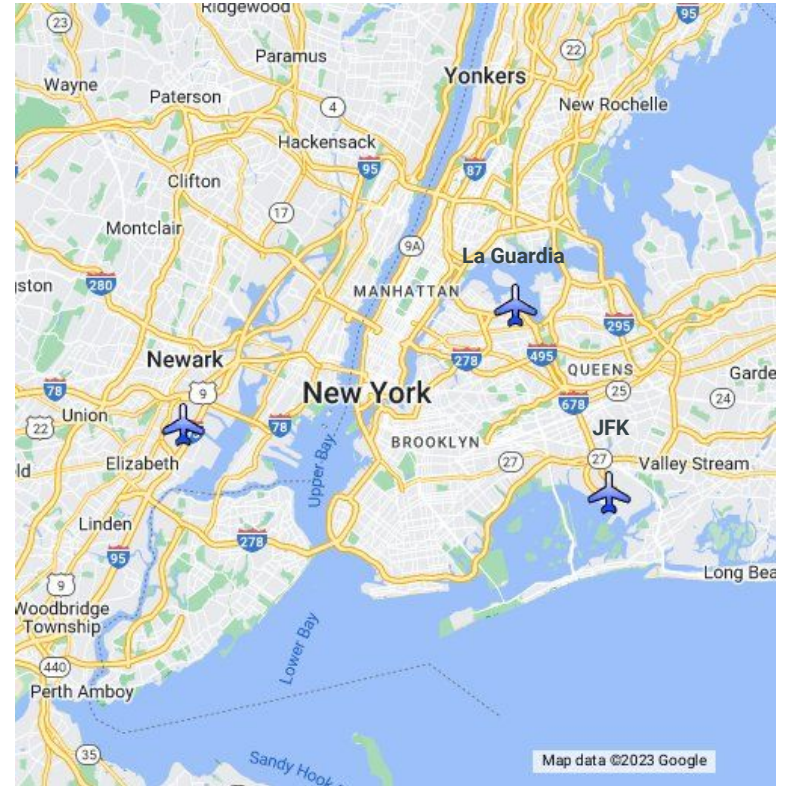
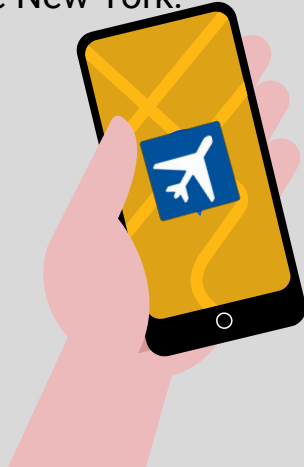
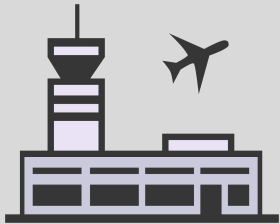


# Création de variables

03

## Aéroport

Création d'une variable **"Course Aéroport"**, pour indiquer les trajets en direction ou en provenance des 3 aéroports de New York.



# Variables de temps

**jour** : jour relatif au trajet ;  
**mois** : mois relatif au trajet ;  
**année** : année relative au trajet ;  
**saison** : saison relative au trajet.

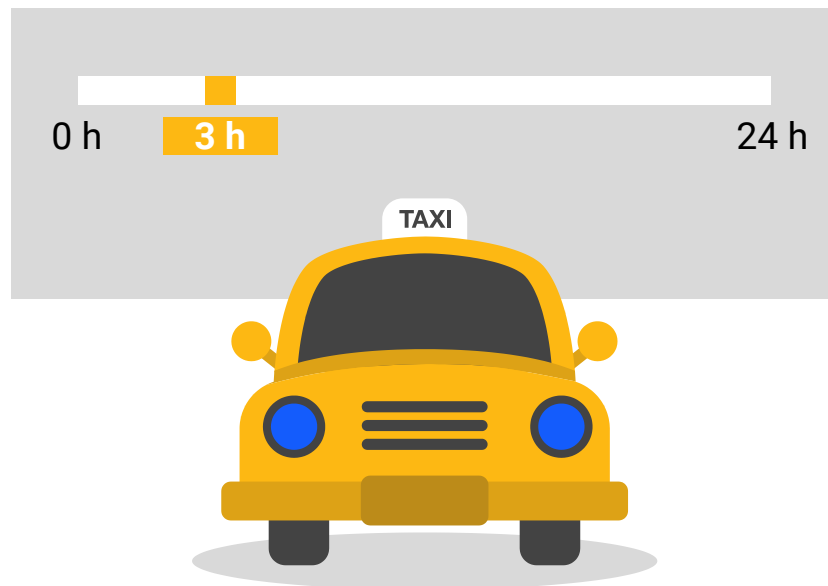


## Heure

L'heure à laquelle le trajet en taxi a commencé.



M	T	W	T	F	S	S
01	02	03	04	05	06	07
08	09	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				



# Statistiques descriptives

01

02

03

04

05

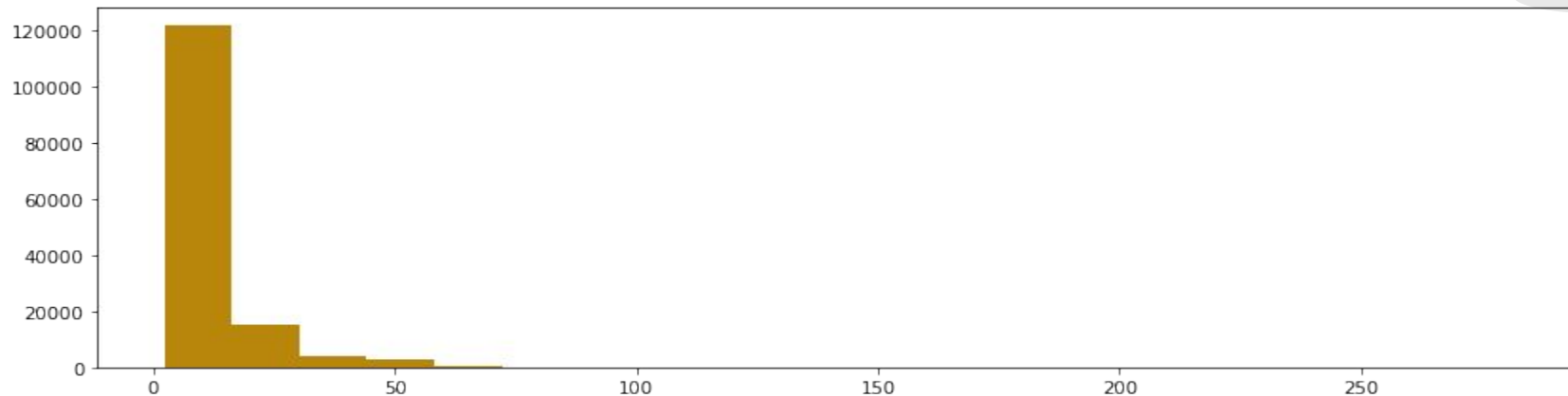


- **Analyse univariée**
- **Analyse bivariée**

# Statistiques descriptives

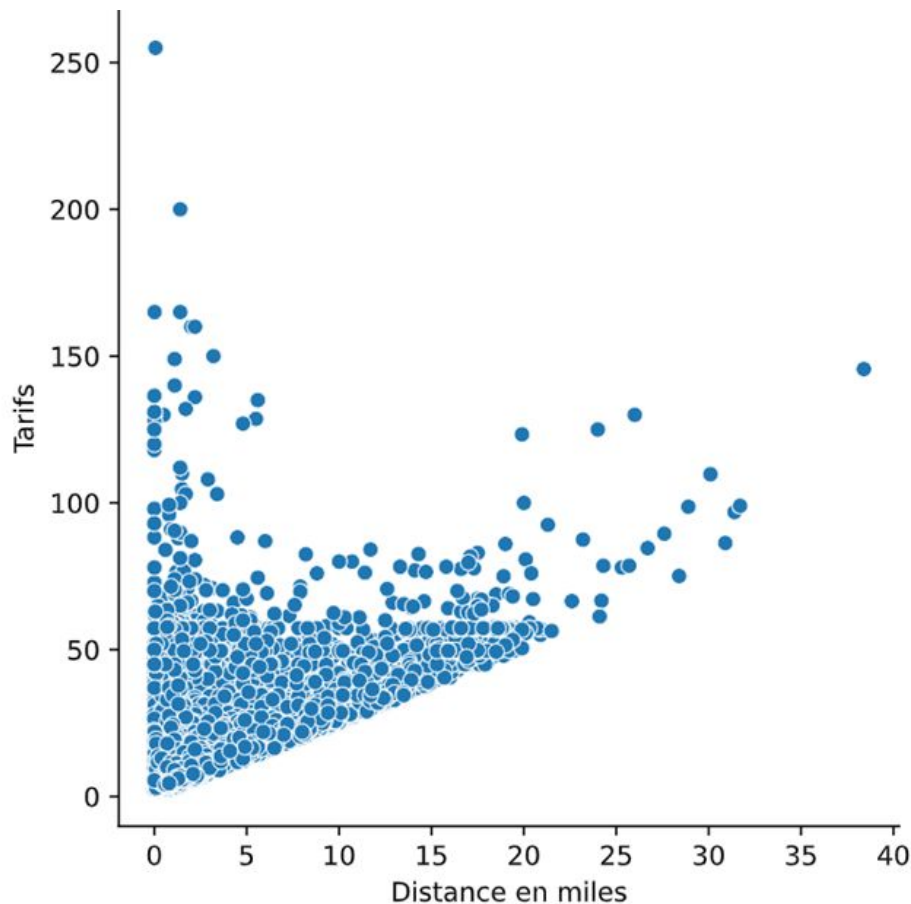


## Tarif des courses en taxi



	Moyenne	Écart-type	Minimum	Médiane	Maximum
Tarif de la course (en \$US)	11,82 \$	10,2	2,5 \$	8,5 \$	500 \$

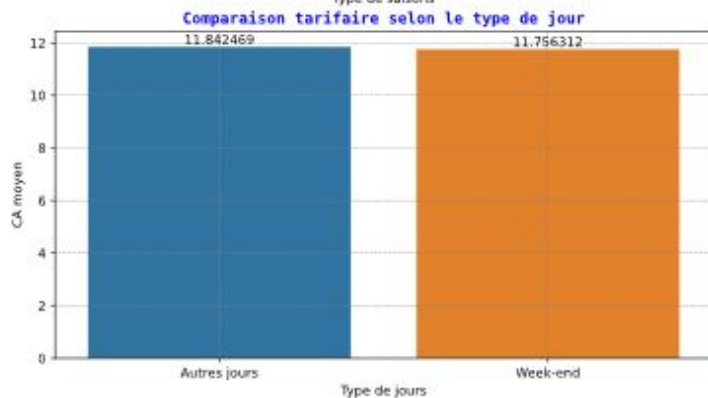
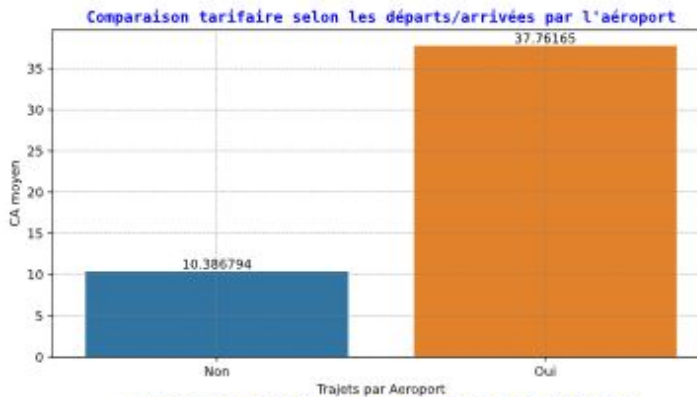
# Évolution du tarif des courses en taxi en fonction de la distance



- trajets de départ ou d'arrivée vers l'aéroport exclus ;
- trajets avec coordonnées de départ et d'arrivée similaires exclus ;
- le tarif de la course évolue en fonction de la distance parcourue.

# Statistiques descriptives

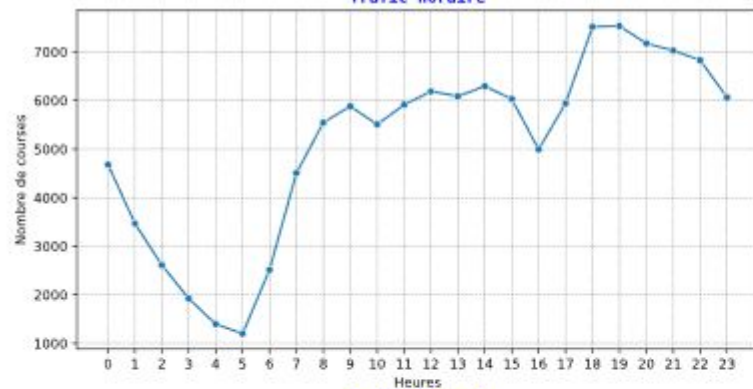
Cible : Tarif des courses en taxi



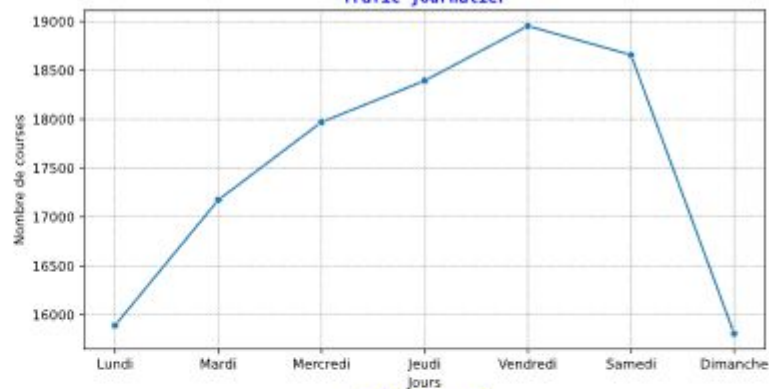
# Informations sur le trafic à New York



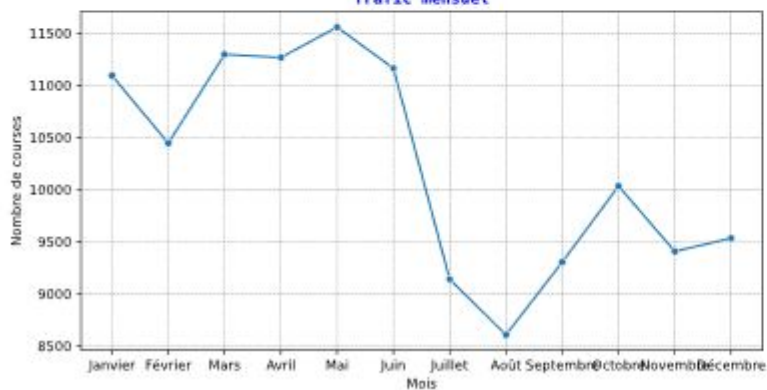
Trafic horaire



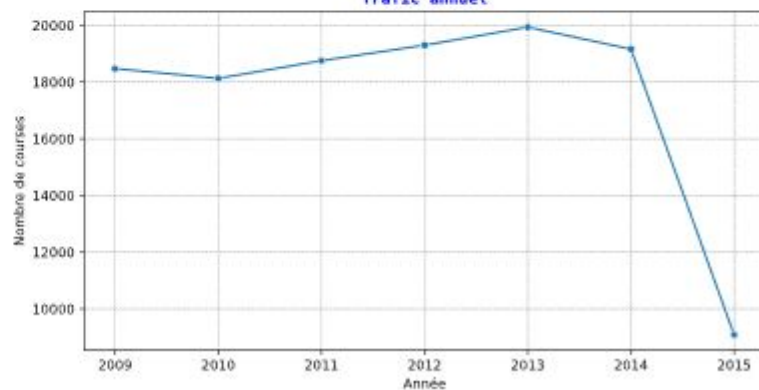
Trafic journalier



Trafic mensuel



Trafic annuel



# Modélisation

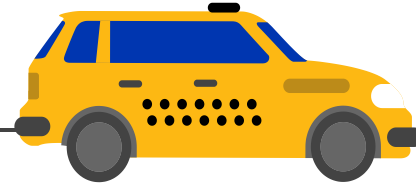
01

02

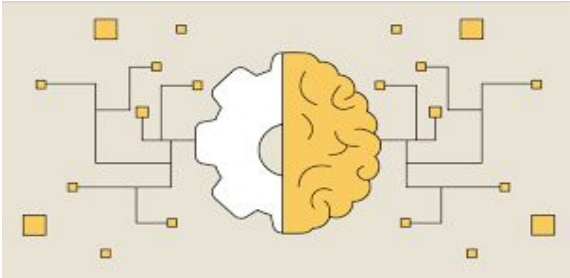
03

04

05



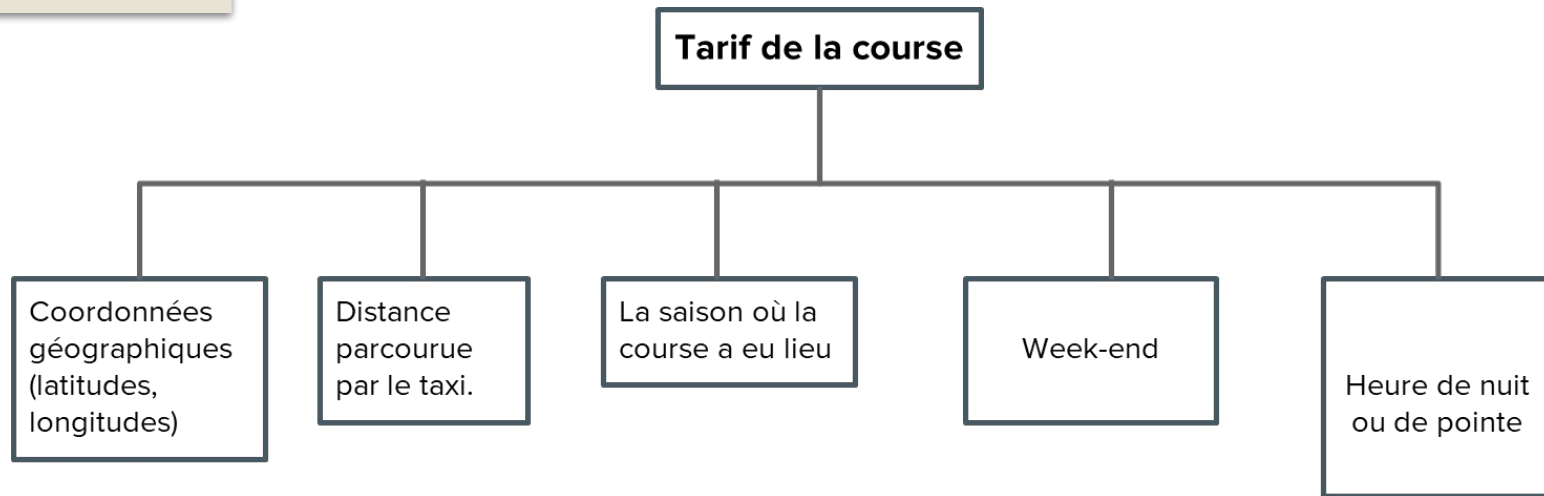
- **Choix des variables**
- **Modèles**
- **Prédiction**
- **Classement Kaggle**







# Modélisation



- ❑ 3 modèles
- ❑ 80 % des données pour l'entraînement
- ❑ 20 % pour la validation





## Choix du meilleur modèle



	Pouvoir explicatif	Erreur absolue moyenne	Erreur quadratique moyenne
Régression linéaire	50 %	4,01 \$US	7,32 \$US
Random Forest	80 %	2,20 \$US	4,64 \$US
XGBoost	81 %	2,19 \$US	4,56 \$US

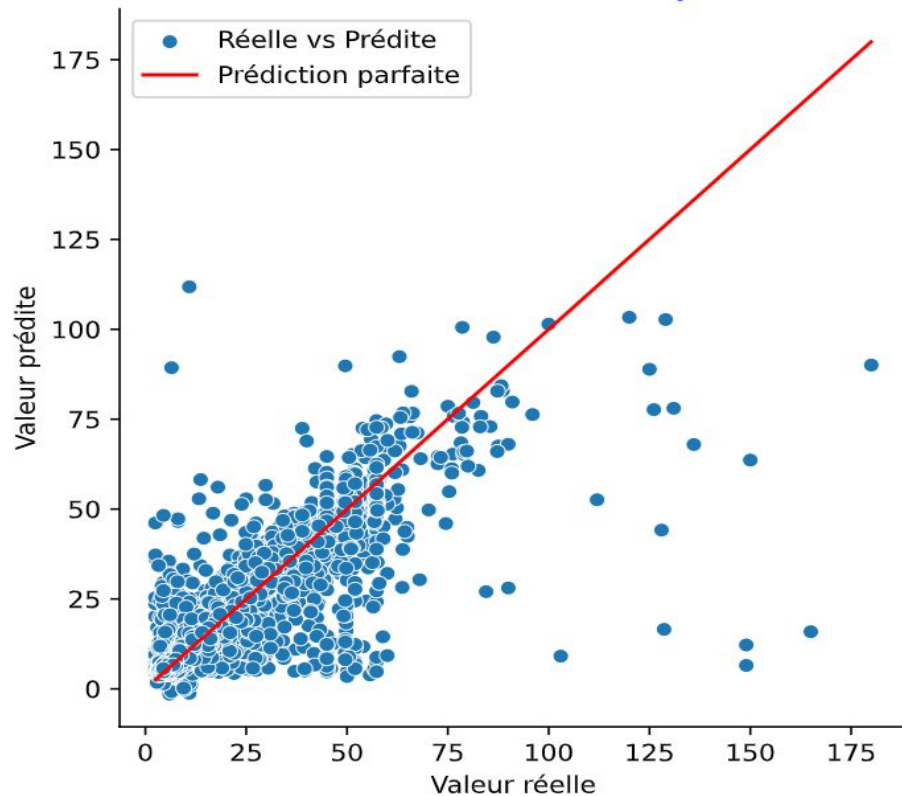
➤ Meilleur modèle : XGBoost




- moins de dispersions ;
- au-dessus de la droite rouge, nous avons les valeurs surestimées ;
- en-dessous de la droite rouge, nous avons les valeurs sous-estimées ;
- prédictions à peu près correctes.



## Prédiction




# Classement sur Kaggle

 Playground Prediction Competition

## New York City Taxi Fare Prediction

Can you predict a rider's taxi fare?

 Google Cloud · 1,483 teams · 4 years ago

[Overview](#) [Data](#) [Code](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Team](#)

[Submissions](#) [Late Submission](#) [...](#)

## Leaderboard

[Raw Data](#)[Refresh](#)

### YOUR RECENT SUBMISSION



Submission.csv

Submitted by Komi Kekeli · Submitted a day ago

Score: 4.69132

1127

ywkim



4.68065

1

5y

1128

UnluckyData



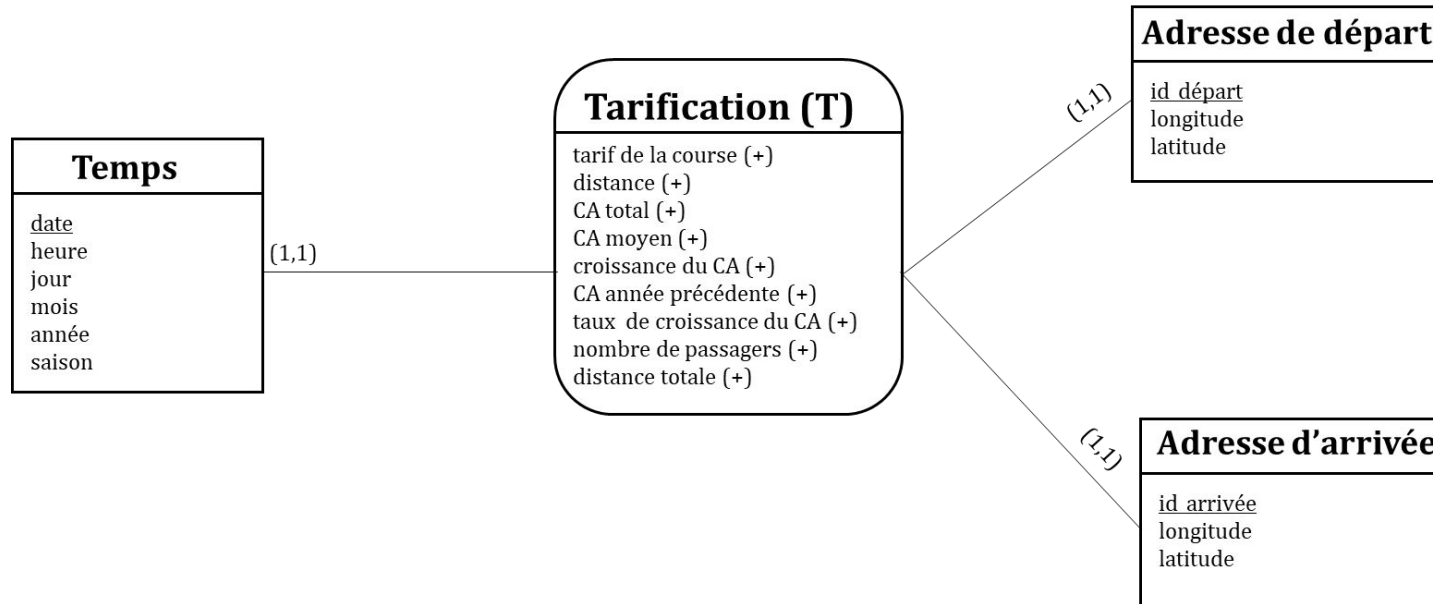
4.70509

2

5y

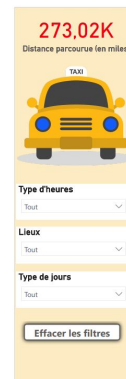
# Schéma factuel

- ❖ Relation factuelle de type transaction

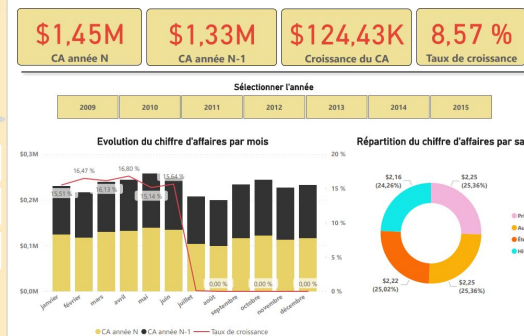


# Outil de restitution

## Analyse du chiffre d'affaires



Analyse du chiffre d'affaires des taxis à New York



# Recommandations



**Service de livraison**



**Service de navettes**



**Système de suivi de flotte**



**Tarifs spéciaux aux clients  
réguliers**

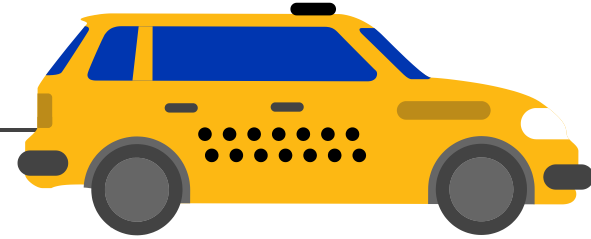
# CONCLUSION

01

02

03

04



## Problématique & étude préalable

- Contexte
- Hypothèses sur le tarif
- Description du jeu de données

## Analyse exploratoire des données

- Traitement de données
- Statistiques descriptives

## Modèles ML










- Modèle (choix des variables)
- Choix du meilleur modèle
- Prédiction
- Classement sur Kaggle

## Restitution

- Schéma factuel
- Outil de visualisation de données (Power BI)
- Recommandations



# Répartition de la charge de travail

	Documentation	Traitement de données	Statistiques descriptives	Modèles ML	Outil de restitution
Amine					
Komi					
Racky					
Reda	