Challenge

Vous travaillerez sur un sous-ensemble des données du challenge Kaggle ECMLPKDD 15: Taxi Trip Time Prediction (II) (voir https://www.kaggle.com/c/pkdd-15-taxi-trip-time-prediction-ii). Les données disponibles sur le site contiennent les variables : `CALL_TYPE, ORIGIN_CALL, ORIGIN_STAND, TAXI_ID, TIMESTAMP, DAYTYPE, MISSING_DATA, POLYLINE`

Sur les données de départ, les transformations suivantes ont été effectuées :

1. les variables `ORIGIN_CALL, ORIGIN_STAND, TAXI_ID` ont été retirées
2. à partir des variables `CALL_TYPE, TIMESTAMP, DAYTYPE, MISSING_DATA, POLYLINE`, les variables suivantes ont été créées
   - `wday` indique le type du jour (entre 0 et 6, le 0 correspond au lundi) de prise en charge
   - `hour` indique l'heure de prise en charge
   - `d_st` indique la distance entre le lieu de prise et le centre-ville (latitude =-8.615223, longitude =41.157819)
   - `heading` indique la direction (par un angle) prise au début de la prise en charge
   - `xs`, `ys` indique la position de la prise en charge (latitude, longitude)
   - `xe`, `ye` indique la position 15sec après la prise en charge (latitude, longitude)
   - `len` indique la longueur du trajet (le nombre de coordonnées GPS enregistrées pendant le trajet, avec un enregistrement toutes les 15sec)
3. un sous-ensemble d'apprentissage `train.csv`, contenant 200000 données a été créé
4. un sous-ensemble de test `test.csv`, contenant 100 données a été créé.

Voici les principales étapes :

1. Lire les détails sur les données sur https://www.kaggle.com/c/pkdd-15-taxi-trip-time-prediction-ii.
2. Expliquer pourquoi il y a de la censure dans le jeu de données `train.csv`. Créer un colonne correspondant au temps de trajet et une à l'indicatrice de censure.
3. Explorer les jeux de données (histogrammes, barplots, etc).
4. Couper votre jeu de données d'apprentissage en 2 `train_train.csv` et `train_test.csv` (90%, 10%, vous pouvez recommencer plusieurs fois le découpage pour éviter les aléas dus au choix des individus) pour pouvoir tester les différents modèles que vous allez construire.
5. Les modèles seront construits sur `train_train.csv` et testés sur `train_test.csv`. Proposer des modèles pour expliquer la distribution du temps de trajet à partir des covariables (toutes les transformations de covariables sont permises et même encouragées....)

6. A partir de ces modèles, proposer une prédiction de la médiane du temps de trajet pour chaque individu du `train_test.csv`. Choisir le meilleur modèle.

7. Estimer dans le modèle choisi avec toutes les données de `train.csv` puis calculer prédiction de la médiane du temps de trajet pour chaque individu du `test.csv`.

Vous m'enverrez un rapport (10 à 15 pages, hors annexes) qui aura la structure

1. Introduction
2. Données
3. Méthodes
4. Résultats
5. Conclusion

et un fichier `results_votrenom.csv` dans lequel vous recopirez les données de `test.csv` et ajouterez une colonne (numéro 12) avec vos prédictions (le séparateur doit être une virgule et l'ordre des colonnes doit être respecté).