

PROJET TITLE

PREDICTION DU TRAFFIC ROUTIER
POUR L'OPTIMISATION DES RÉSEAUX
IOT DE TRANSPORT INTELLIGENT

Presented by:

ARIEL
YAMEO



TABLE OF CONTENTS

01

INTRODUCTION

02

DATA PREPARATION

03

EXPLORATORY ANALYSIS

04

TRAFFIC TREND ANALYSIS

05

MODEL DEVELOPMENT

06

RESULTS & CONCLUSION

INTRODUCTION

De nos jours, l'expansion des réseaux de capteurs IoT dans le secteur du transport permet une collecte continue de données sur l'état du trafic routier. Cependant, dans les zones urbaines, la congestion routière demeure un défi majeur, entraînant une perte de temps, une hausse de la pollution et une diminution de l'efficacité des systèmes de transport.

Dans ce contexte, la prédiction du trafic constitue un levier essentiel pour optimiser la gestion des infrastructures, améliorer la mobilité urbaine et soutenir le développement des systèmes de transport intelligents (ITS – Intelligent Transportation Systems).

Ce projet s'inscrit dans cette dynamique. Il vise à prévoir le volume du trafic et la vitesse des véhicules sur l'autoroute I-94 à partir de données historiques et météorologiques, en exploitant différentes approches d'apprentissage profond (Deep Learning) telles que LSTM, GRU, CNN, Transformer, etc.

DATASET PREPARATION

NOM: Metro_Interstate_Traffic_Volume_with_speed.csv

- Number of lines : 48204
- Number of column: 10
- Missing Value : colonne holidays 48143
- Duplicate Value : 17

#	Column	Dtype
0	traffic_volume	int64
1	holiday	object
2	temp	float64
3	rain_1h	float64
4	snow_1h	float64
5	clouds_all	int64
6	weather_main	object
7	weather_description	object
8	date_time	object

	traffic_volume	holiday	temp	rain_1h	snow_1h	clouds_all	weather_main	weather_description	date_time	speed
0	5545	NaN	288.28	0.0	0.0	40	Clouds	scattered clouds	02-10-2012 09:00	5.0
1	4516	NaN	289.36	0.0	0.0	75	Clouds	broken clouds	02-10-2012 10:00	5.0
2	4767	NaN	289.58	0.0	0.0	90	Clouds	overcast clouds	02-10-2012 11:00	5.0
3	5026	NaN	290.13	0.0	0.0	90	Clouds	overcast clouds	02-10-2012 12:00	5.0
4	4918	NaN	291.14	0.0	0.0	75	Clouds	broken clouds	02-10-2012 13:00	5.0

GESTION DE MISSING_VALUE & DUPLICAT

traitement des missing values

- remplace les nan par no holiday
- creer une autre colonnes ou les jours feries sont marques
1 et les autres 0

traitement des valeurs duplicate

- suppression des valeurs dupliquees

```
## traitement des missing values
data['holiday'] = data['holiday'].fillna('No Holiday')      #remplace les r
data['is_holiday'] = data['holiday'].apply(lambda x: 0 if x == 'No Holiday' else 1)
# drop duplicates
data.drop_duplicates()
print(f"{data.duplicated().sum()} are drop from dataset.")
```

17 are drop from dataset.

	index	Na
0	traffic_volume	0
1	holiday	0
2	temp	0
3	rain_1h	0
4	snow_1h	0
5	clouds_all	0
6	weather_main	0
7	weather_description	0
8	date_time	0
9	speed	0
10	is_holiday	0

FEATURE ENGINEERING

À cette étape, nous allons extraire des caractéristiques utiles à partir de la colonne date_time, convertir les variables catégorielles en format numérique, et normaliser ou les caractéristiques numériques

Formatage des dates + extraction des variables temporelles (annee, jour, mois, jour de la semaine,heure).

		traffic_volume	holiday	temp	rain_1h	snow_1h	clouds_all	weather_main	weather_description	speed	year	month	day	day_of_week	hour
	date_time														
	2012-10-02 09:00:00	5545	NaN	288.28	0.0	0.0	40	Clouds	scattered clouds	5.0	2012	10	2	1	9
	2012-10-02 10:00:00	4516	NaN	289.36	0.0	0.0	75	Clouds	broken clouds	5.0	2012	10	2	1	10
	2012-10-02 11:00:00	4767	NaN	289.58	0.0	0.0	90	Clouds	overcast clouds	5.0	2012	10	2	1	11
	2012-10-02 12:00:00	5026	NaN	290.13	0.0	0.0	90	Clouds	overcast clouds	5.0	2012	10	2	1	12
	2012-10-02 13:00:00	4918	NaN	291.14	0.0	0.0	75	Clouds	broken clouds	5.0	2012	10	2	1	13

ONE-HOT ENCODING

ici, on va préparer le data en transformant les colonnes catégorielles en numeric pour pouvoir l'utiliser dans les différents modèles

```
#Converting Categorical values to Numerical using one-hot-encoding
categorical_cols = ['holiday', 'weather_main', 'weather_description']

# Vérifie que ces colonnes existent avant de les encoder
categorical_cols = [col for col in categorical_cols if col in data.columns]

data = pd.get_dummies(data, columns=categorical_cols, drop_first=True)
```

EXPLORATORY ANALYSIS

```
traffic_volume'].describe()
```

```
48204.000000  
3259.818355  
1986.860670  
0.000000  
1193.000000  
3380.000000  
4933.000000  
7280.000000
```

```
data['speed'].describe()
```

```
count      48204.000000  
mean       14.392969  
std        20.348985  
min        5.000000  
25%        5.000000  
50%        5.000000  
75%        5.000000  
max        100.000000  
Name: speed, dtype: float64
```

Le trafic routier varie fortement : parfois la route est presque vide, parfois elle est très chargée. L'écart-type élevé montre que cette variation est importante, probablement à cause des heures de pointe, de la météo ou des jours particuliers. La moyenne proche de la médiane indique que la plupart du temps le trafic est faible ou moyen, avec quelques pics de forte affluence. En résumé, le volume de trafic dépend fortement du moment de la journée et des conditions extérieures.

LA VITESSE DES VÉHICULES VARIE BEAUCOUP SELON LES CONDITIONS : PARFOIS LA CIRCULATION EST FLUIDE, PARFOIS RALENTIE, NOTAMMENT AUX HEURES DE POINTE OU EN CAS DE CONGESTION. LA DISTRIBUTION DES VITESSES EST RELATIVEMENT SYMÉTRIQUE, AVEC QUELQUES VALEURS EXTRÊMES. EN GÉNÉRAL, LA VITESSE MOYENNE RESTE STABLE, MAIS ELLE PEUT CHUTER LORS DES PÉRIODES DE FORTE AFFLUENCE OU D'INCIDENTS.

WHITE NOISE & RANDOM WALK

WHITE NOISE



“Le test de Ljung-Box appliqué aux résidus du modèle de persistance donne une p-valeur de 0,01 (< 0,05). On rejette donc l’hypothèse nulle : les résidus ne sont pas du bruit blanc, ce qui indique qu’il existe des dépendances temporelles dans la série.”

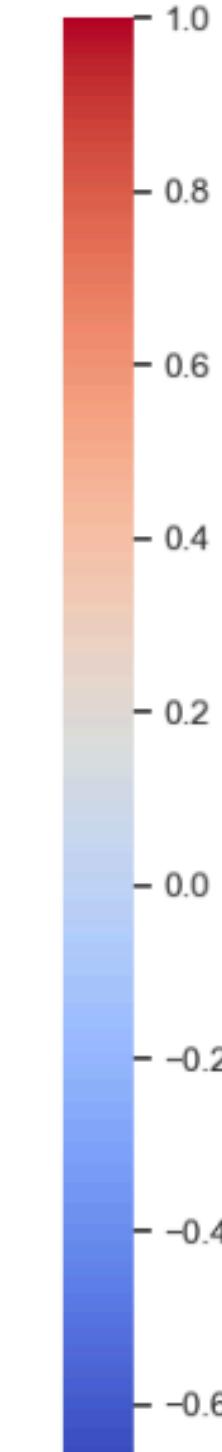
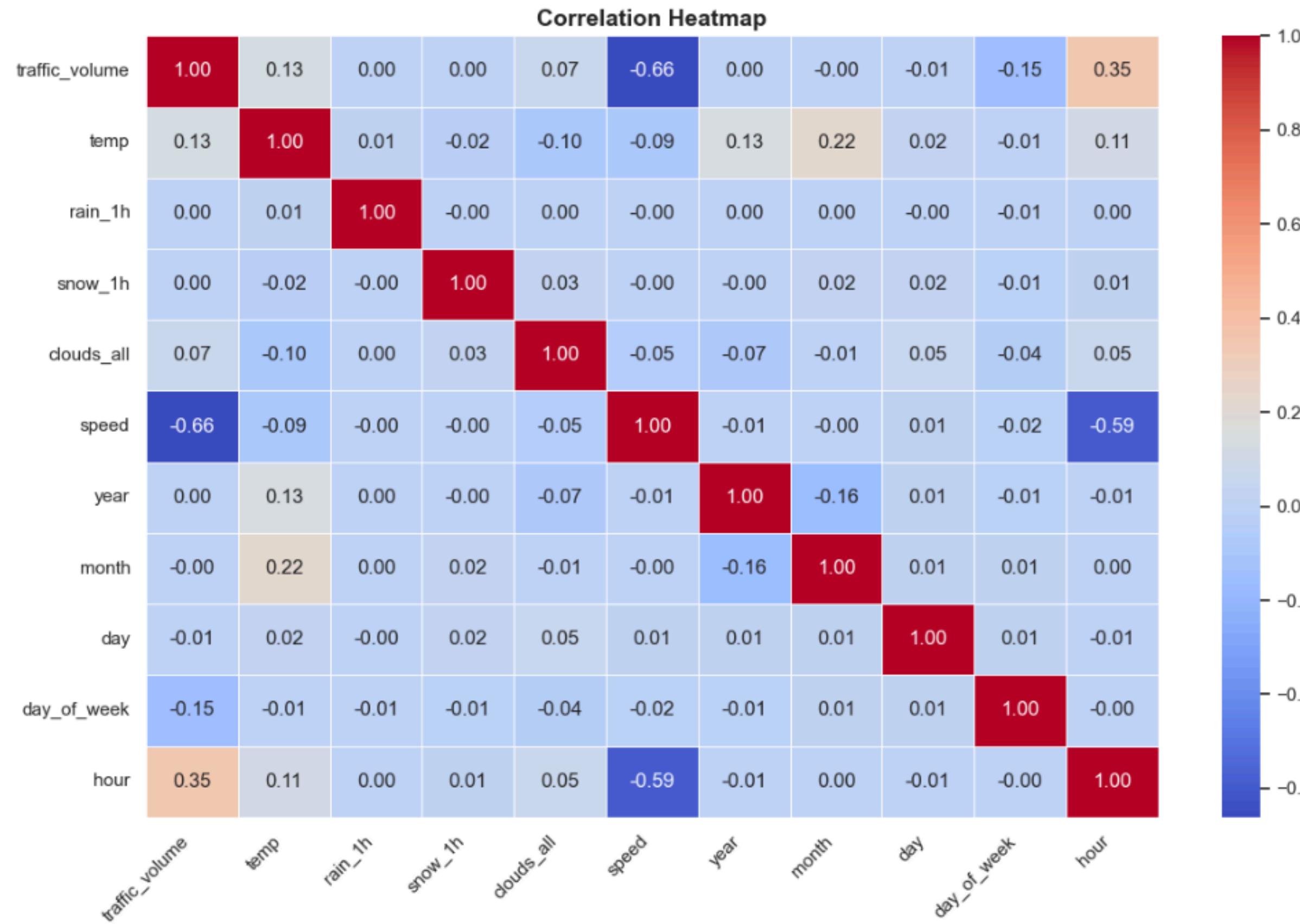
RANDOM WALK



ADF Statistic: -28.016624319503496 p-value: 0.0

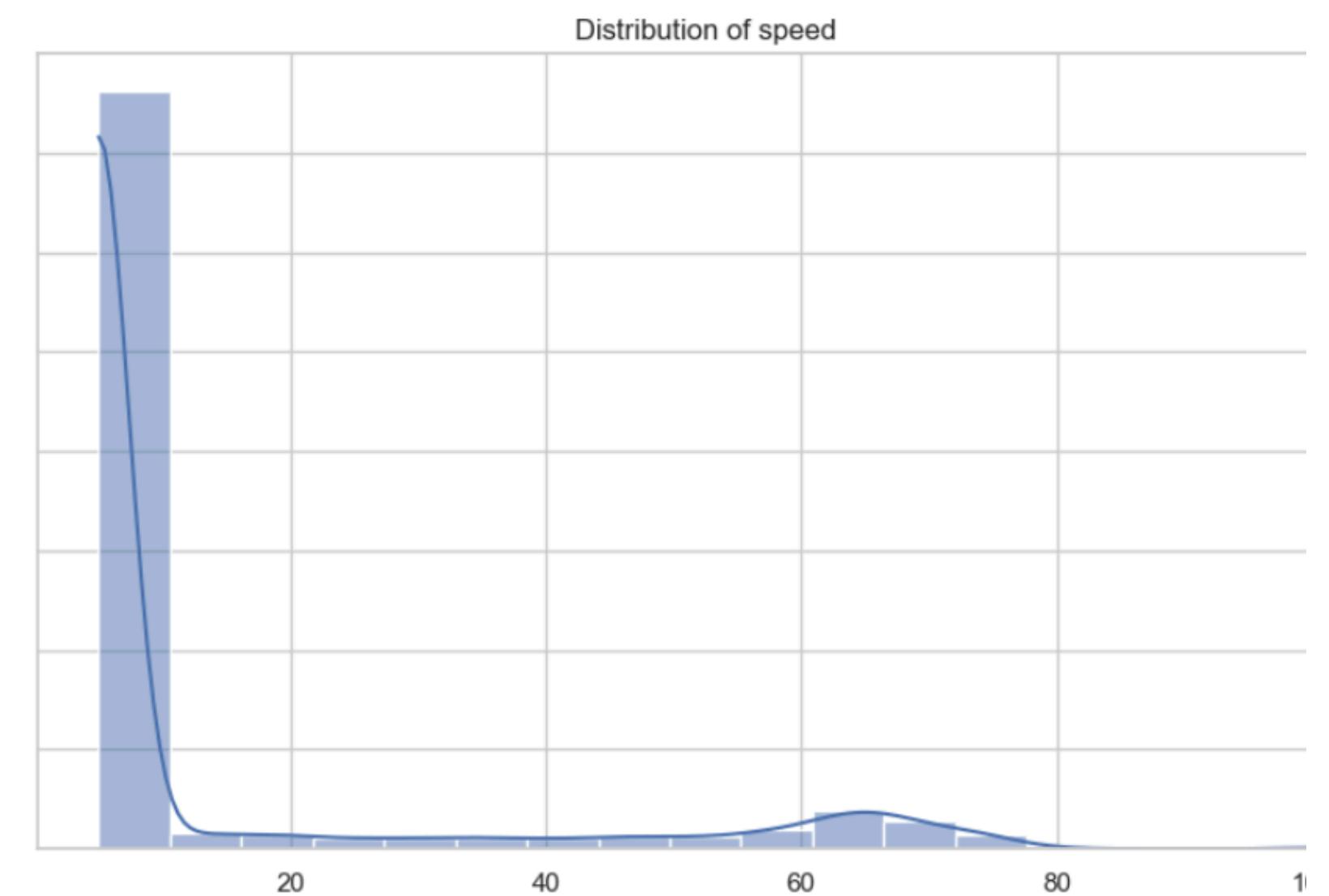
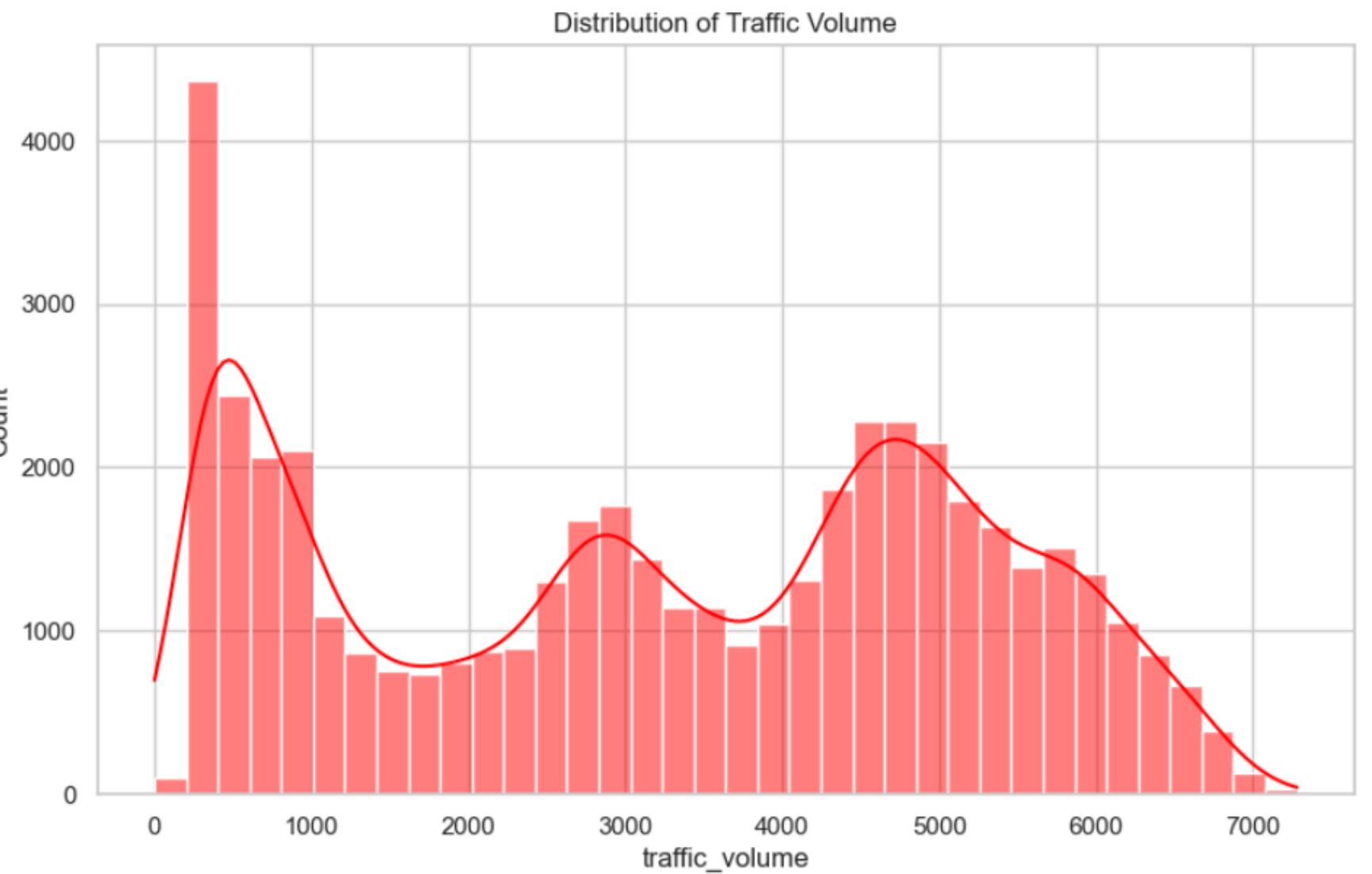
“Le test de Dickey-Fuller (ADF) sur la série du trafic montre une p-valeur de 0,12 (> 0,05). On ne peut donc pas rejeter l’hypothèse nulle : la série est probablement non stationnaire et présente un comportement proche d’un Random Walk.”

HEATMAP



En résumé, les facteurs temporels (heure, jour de la semaine) et la vitesse des véhicules apparaissent comme les variables les plus déterminantes pour expliquer les variations du volume de trafic, tandis que les conditions météorologiques ont un effet limité.

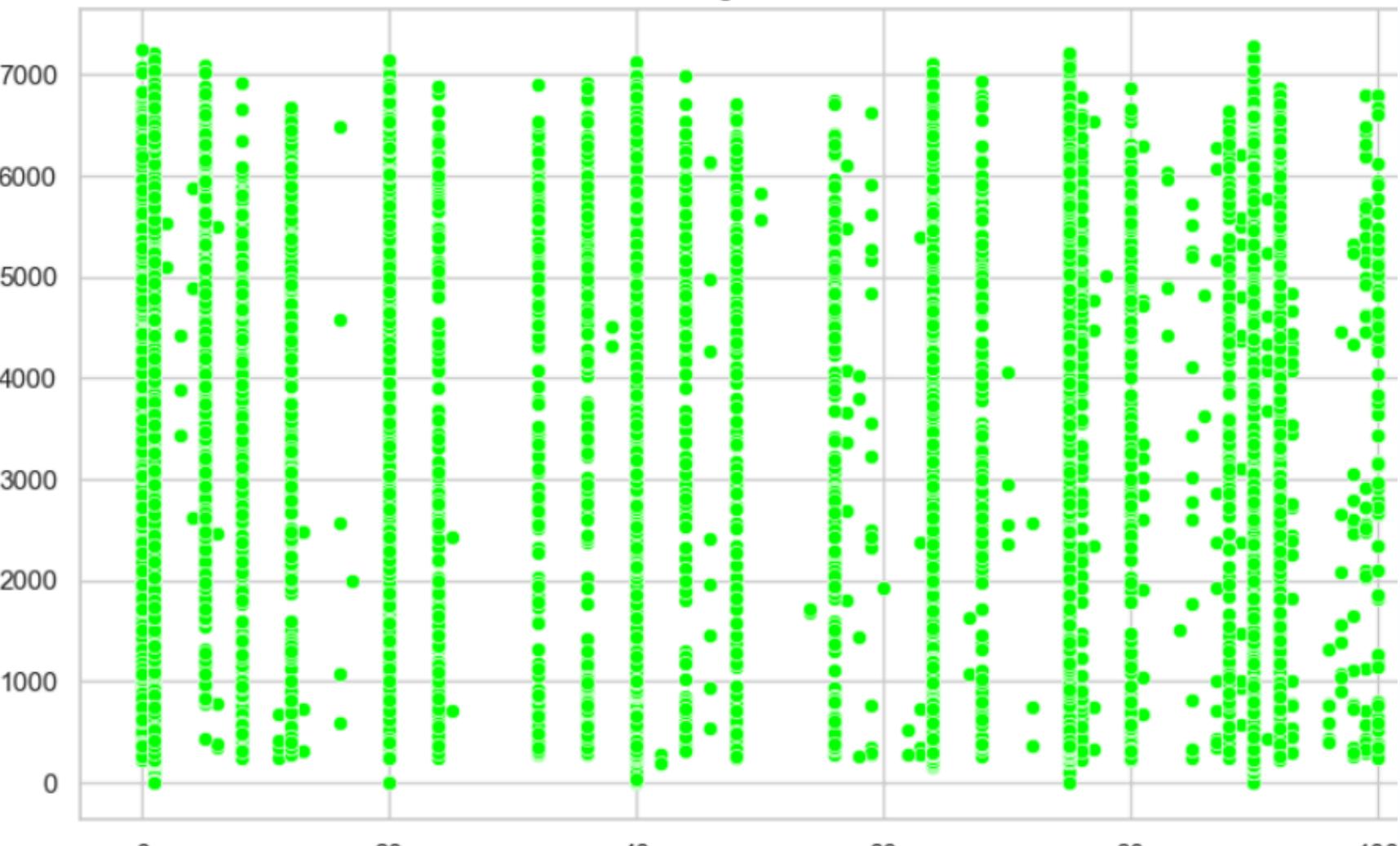
EXPLORATORY ANALYSIS



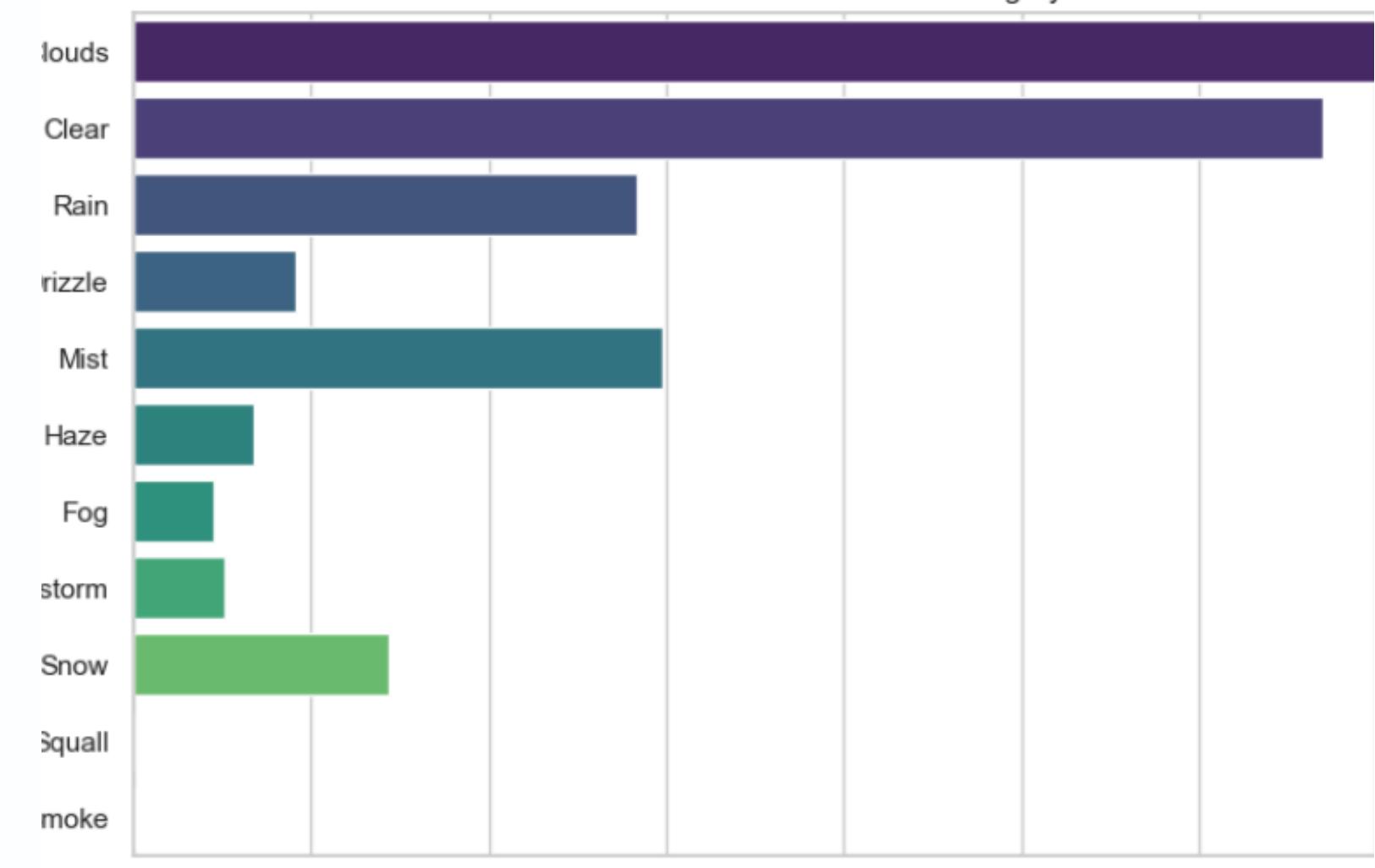
Ajouter des lignes dans le corps du texte

EXPLORATORY ANALYSIS

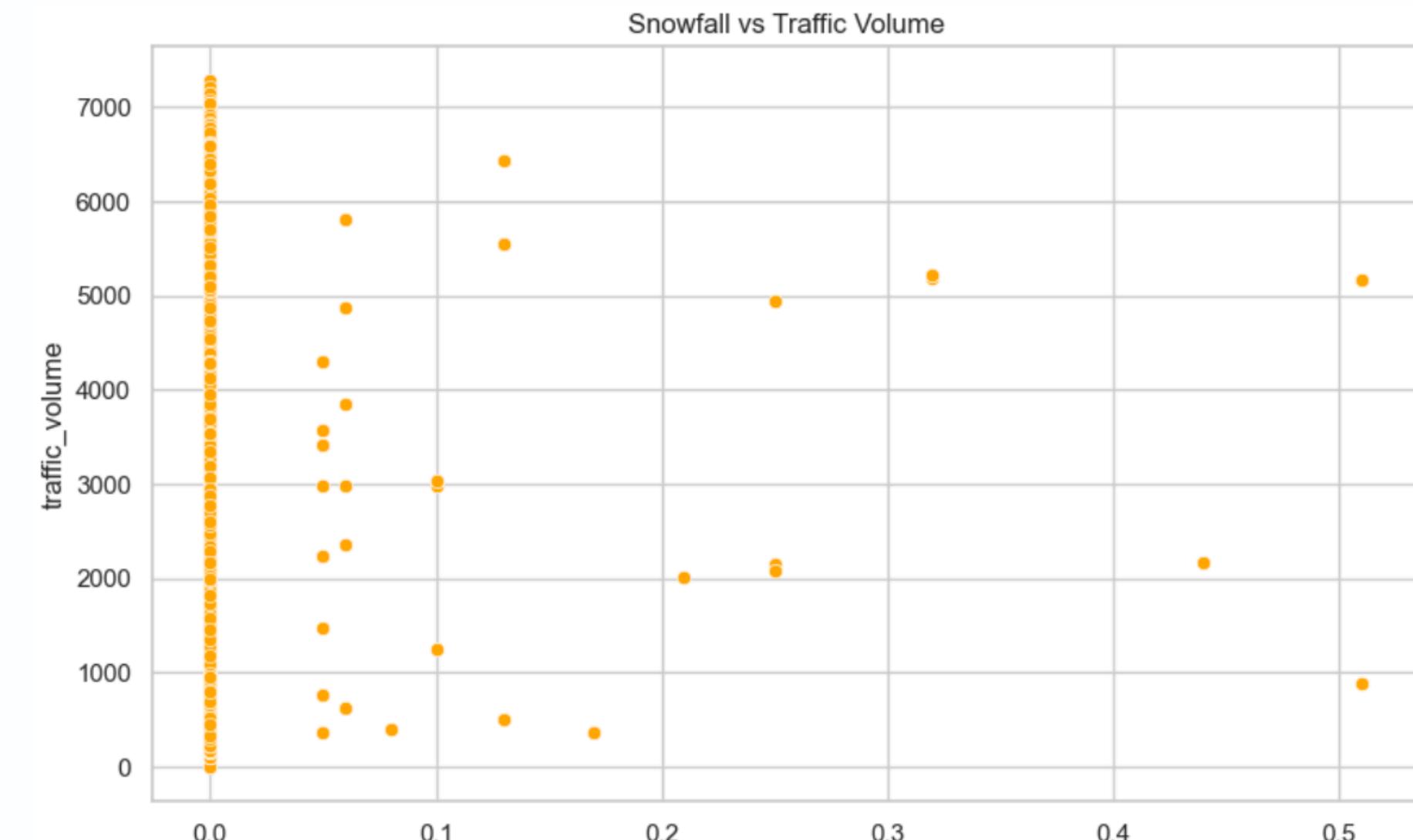
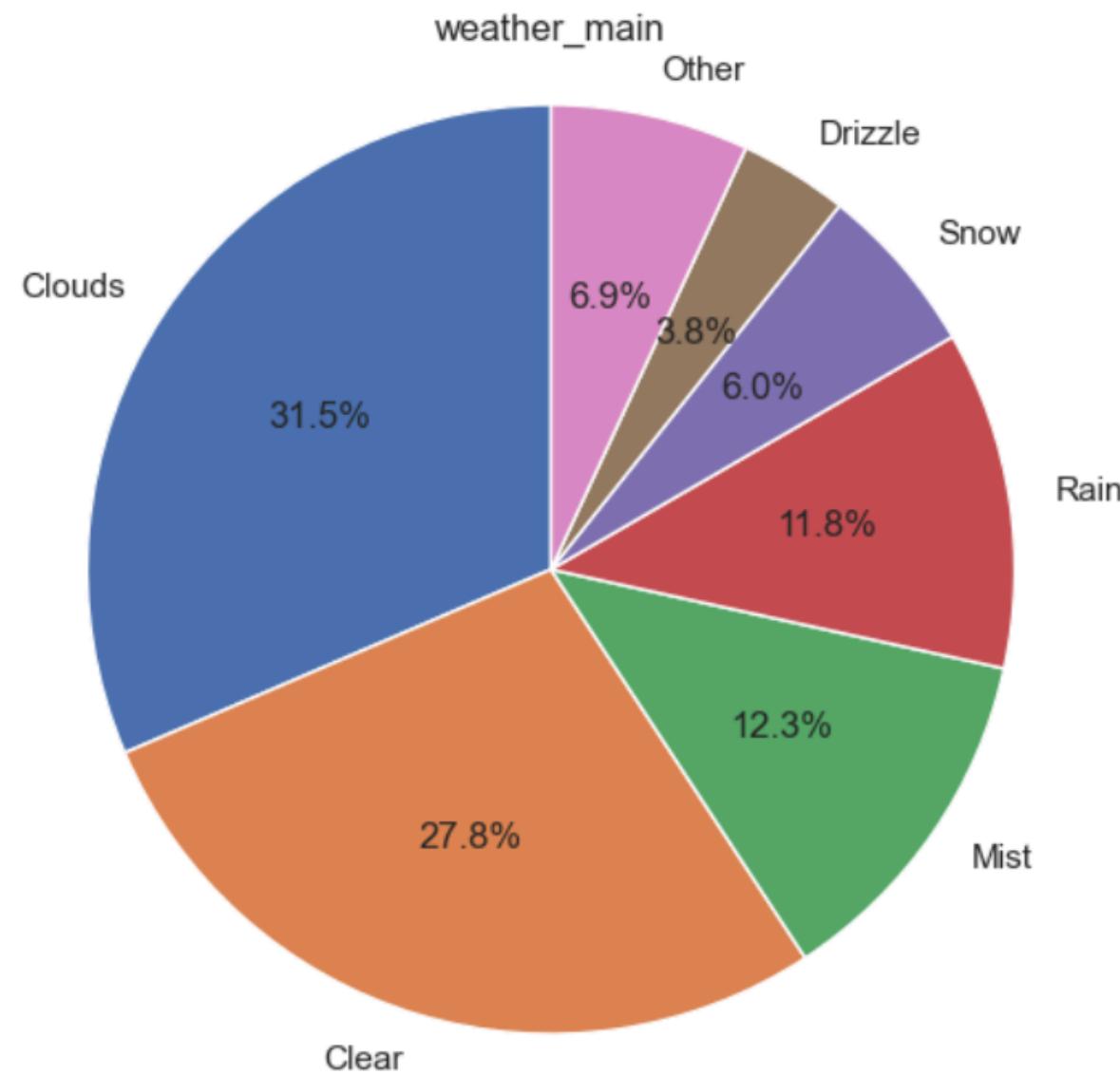
Cloud Coverage vs Traffic Volume



Count of Each Weather Category



EXPLORATORY ANALYSIS



TRAFFIC TREND ANALYSIS

TIME INDICATOR

L'un des facteurs possibles d'un trafic élevé est le temps.

Il peut y avoir plus de personnes sur la route à certains mois, certains jours de la semaine ou à certaines heures de la journée.

Nous allons donc examiner quelques graphiques linéaires montrant comment le volume du trafic varie en fonction des éléments suivants:

Le mois

Le jour de la semaine

L'heure de la journée

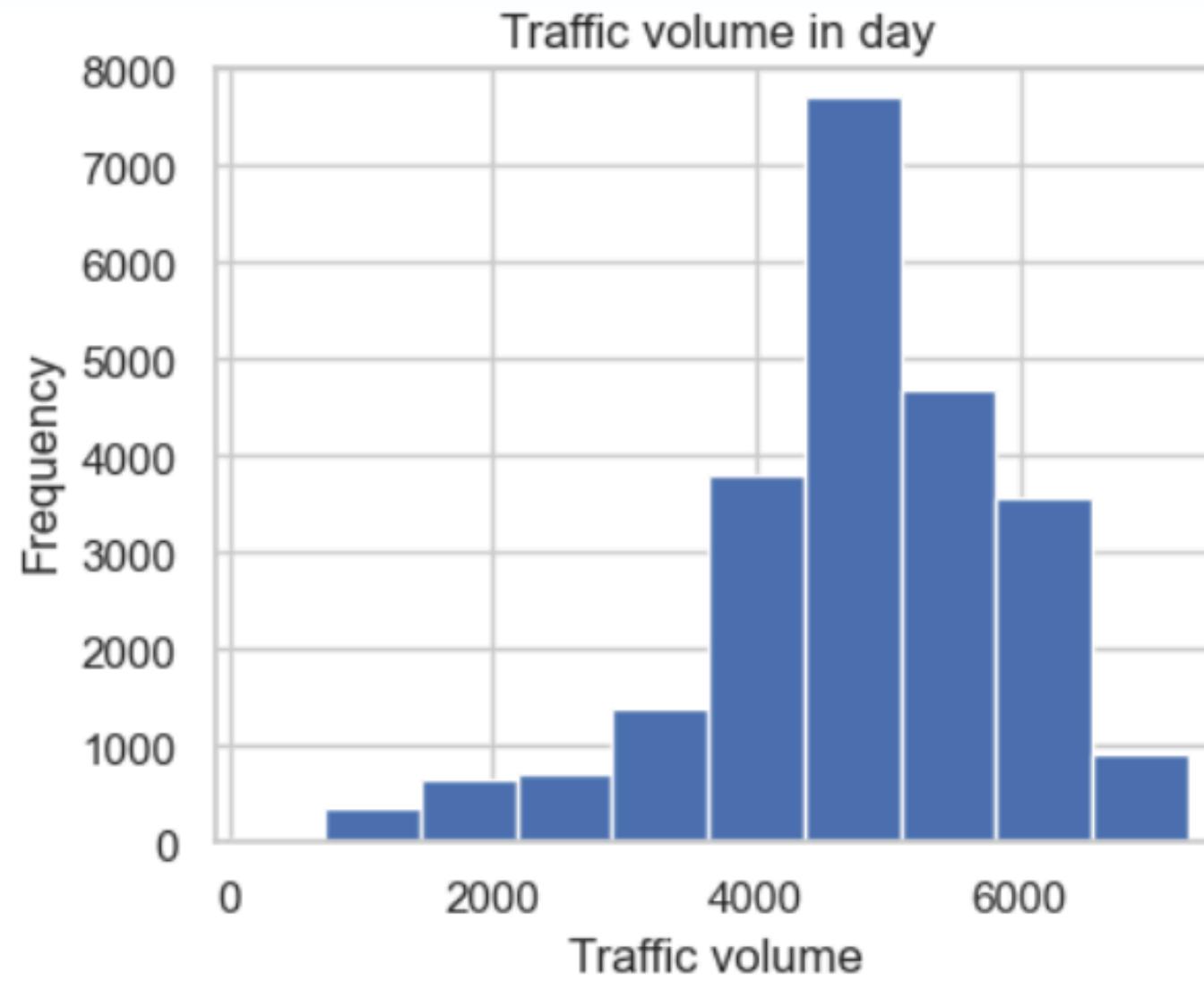
Weather indicators

Un autre indicateur possible d'un trafic intense est la météo.

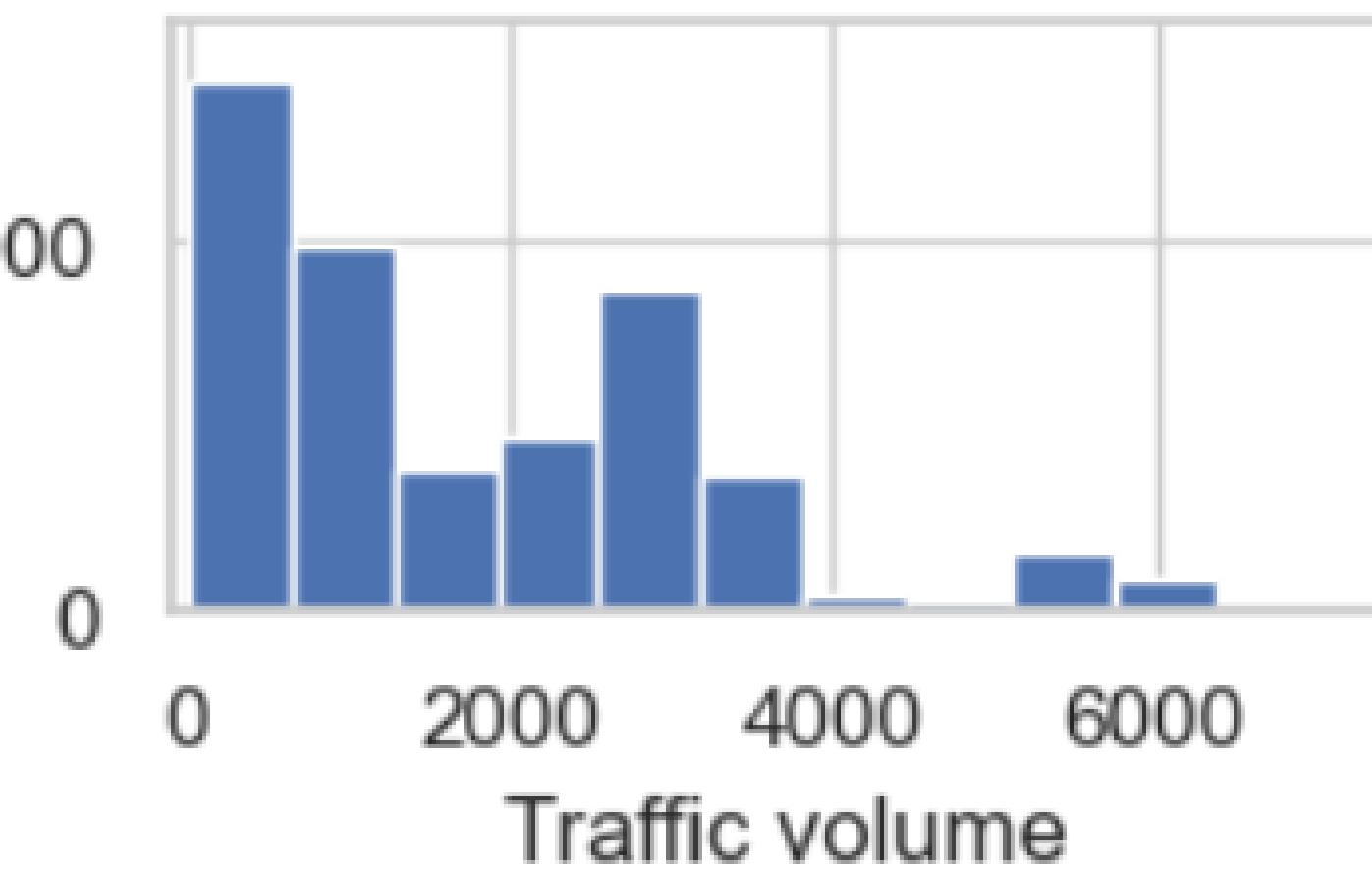
Le jeu de données nous fournit quelques colonnes utiles concernant la météo : temp, rain_1h, snow_1h, clouds_all, weather_main, weather_description.

Certaines de ces colonnes sont numériques, nous allons donc commencer par examiner leurs valeurs de corrélation avec le volume de trafic.

TRAFFIC DAY / NIGHT



Traffic volume in night

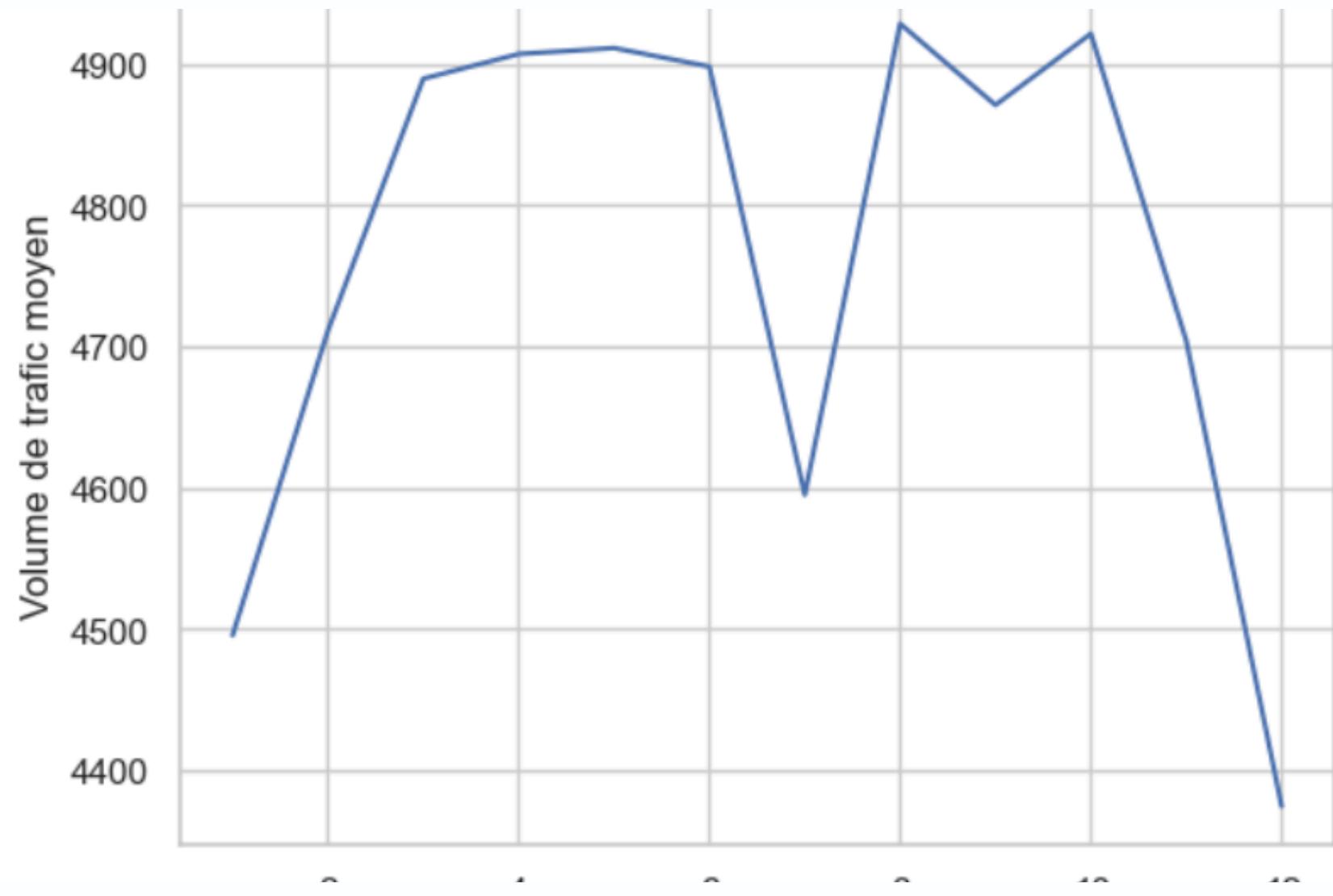


Le graphique des données diurnes est asymétrique vers la gauche ⇒ la plupart du temps, le volume de trafic pendant la journée est élevé. Dans 75 % des cas, le volume de trafic est supérieur à 4 252.

Pendant la nuit, la distribution est asymétrique vers la droite ⇒ la majorité du temps, le volume de trafic est faible. Dans 75 % des cas, le volume de trafic est inférieur à 2 819.

Donc le traffic est plus impressionnant le jour que la nuit

TRAFFIC BY MONTH



Les données montrent que les mois de mars à juin enregistrent un plus grand nombre de véhicules, tout comme la période d'août à octobre. Cela pourrait s'expliquer par le facteur météorologique : ce sont des périodes où le temps est agréable, et les gens sortent plus souvent que pendant les autres mois.

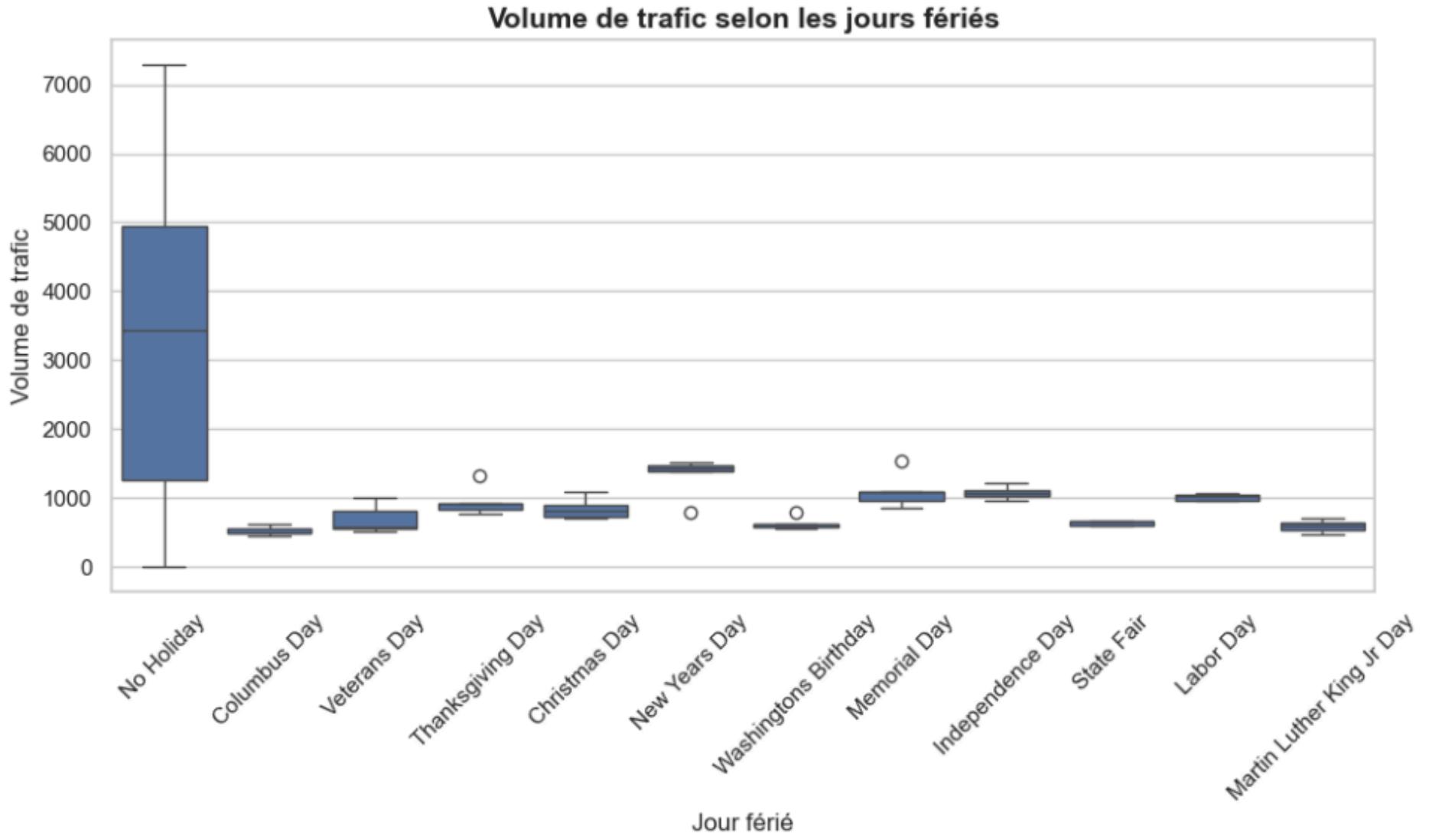
Il est possible que les mois de janvier, juillet et décembre présentent un nombre beaucoup plus faible de véhicules, car ils sont fortement influencés par les conditions climatiques:

Chaleur en été (juillet)

Neige en hiver (décembre – janvier)

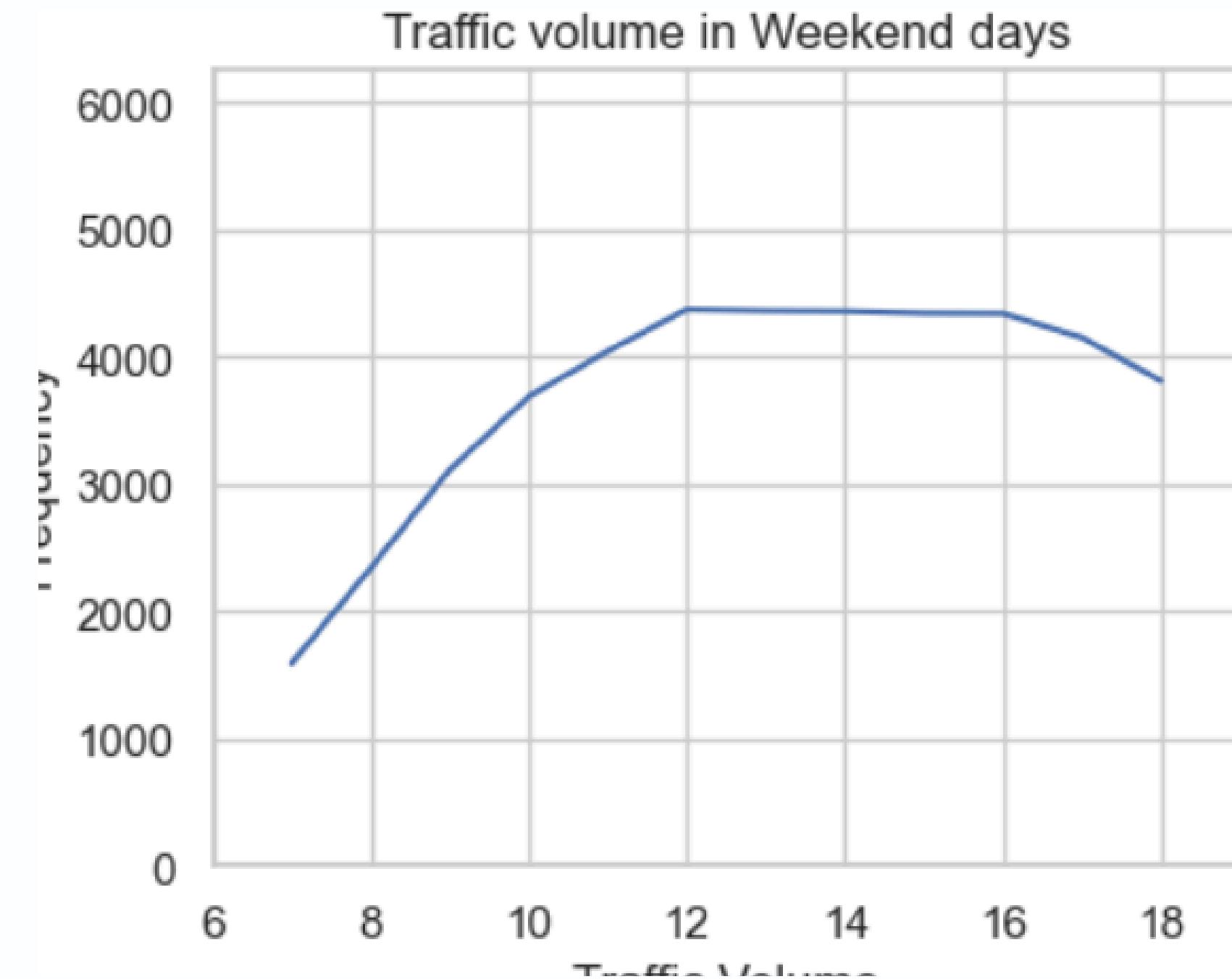
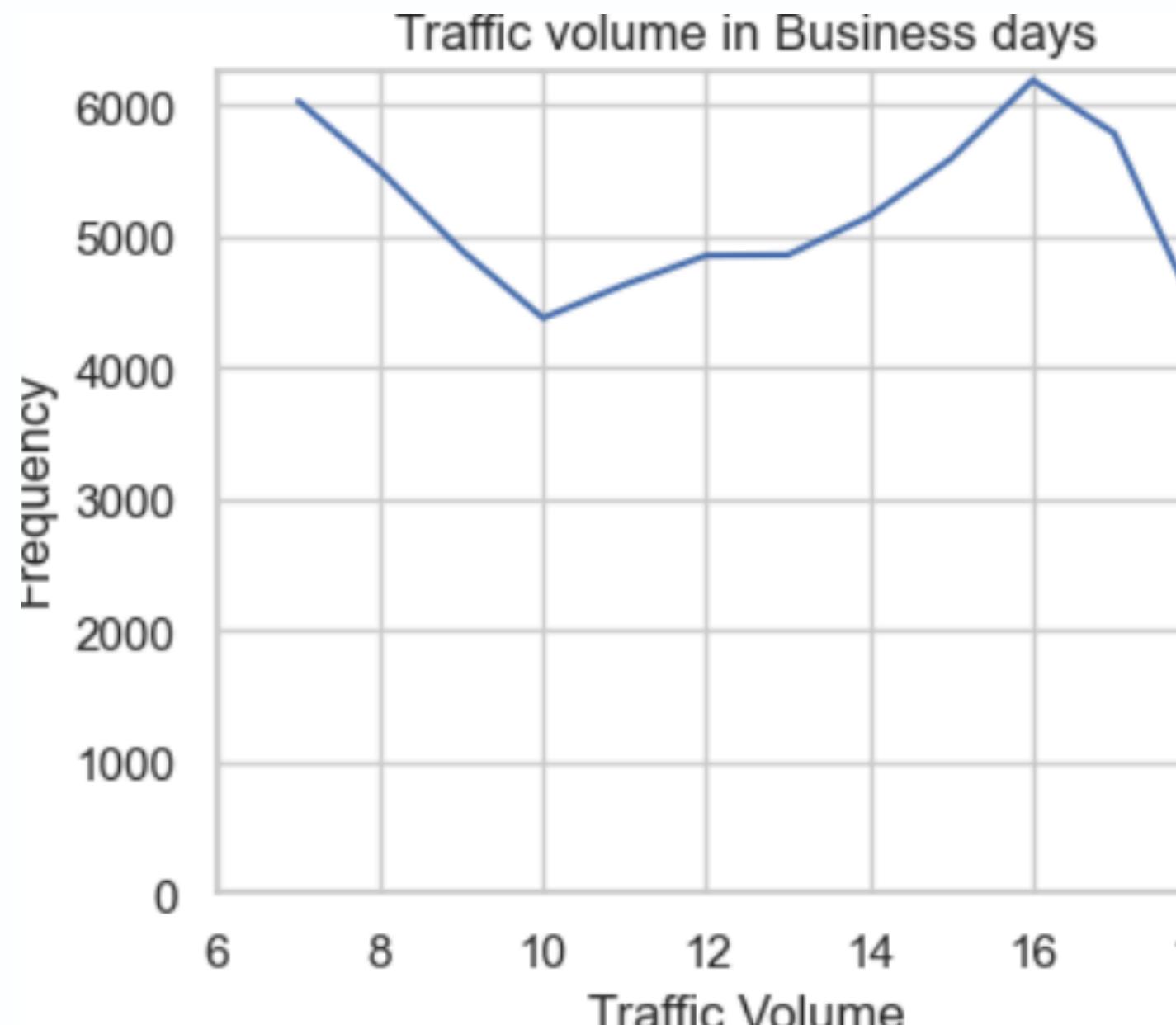
NB: on a pris le jour car c'est le facteur avec plus de densité que la nuit

TRAFFIC BY HOLIDAY



Le boxplot met en évidence les différences de volume de trafic entre les jours fériés et les jours ordinaires. On observe globalement que le volume de trafic est plus faible pendant les jours fériés, ce qui s'explique par la réduction des déplacements professionnels. Cette tendance suggère que la variable **holiday** peut être un facteur explicatif important dans la modélisation du trafic.

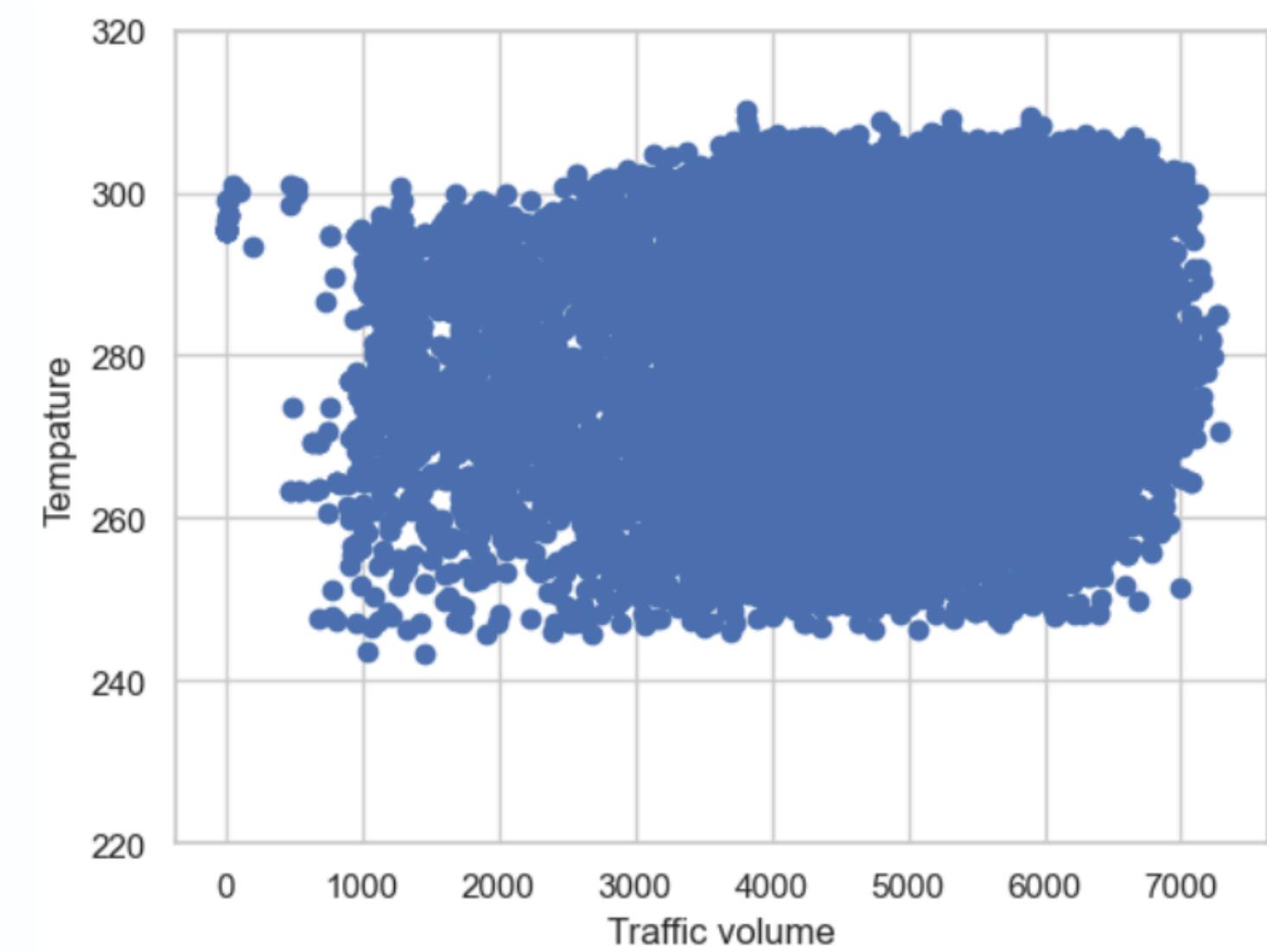
TRAFFIC BUSINESS DAY / WEEKEND DAY



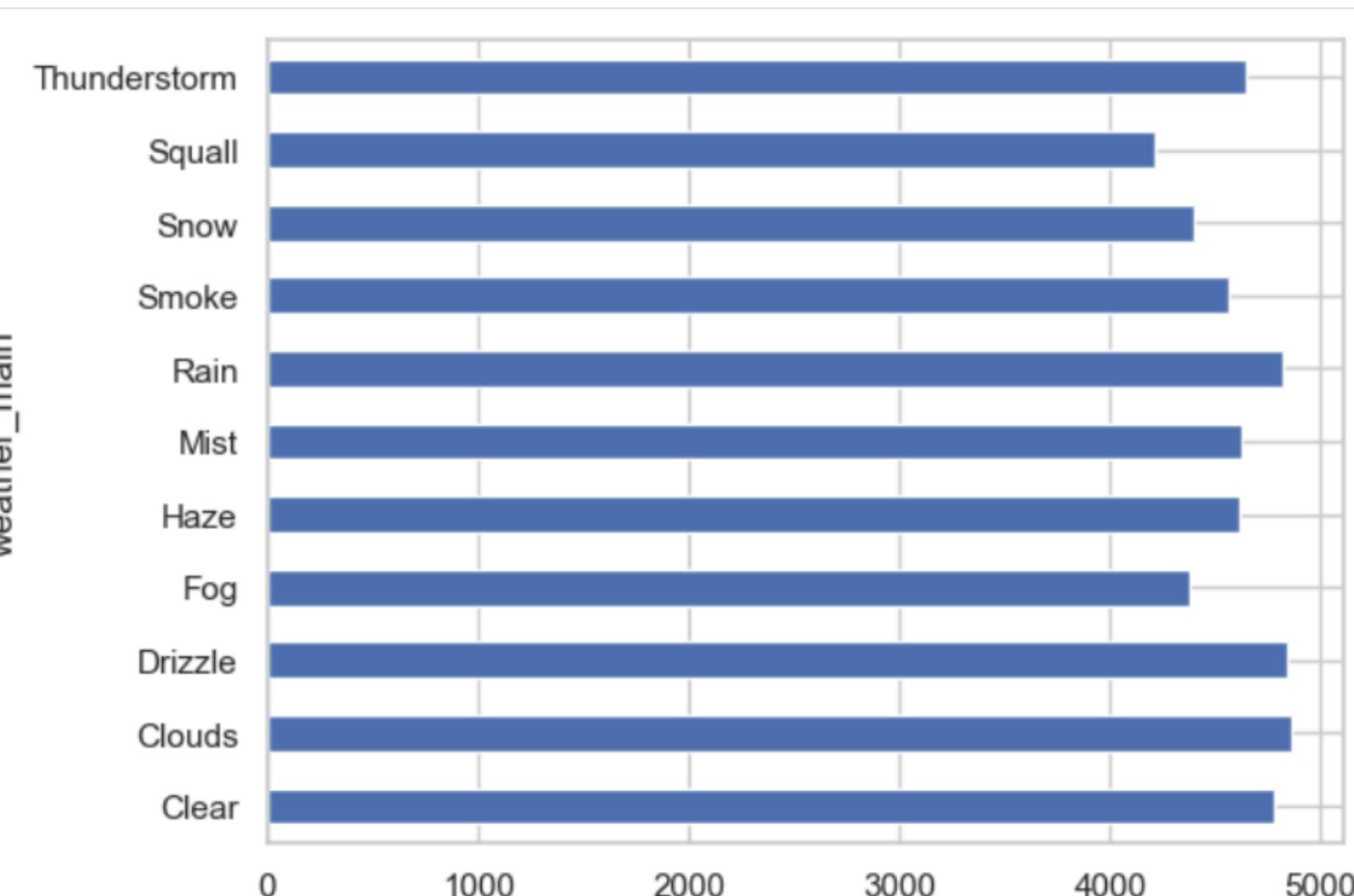
On peut clairement constater que les périodes de 7 à 8 heures du matin et de 16 à 17 heures de l'après-midi correspondent aux heures de pointe, le volume de trafic dépassant 5 000 véhicules pendant les jours ouvrables. En revanche, ces périodes correspondent aux volumes les plus faibles pendant le week-end.

Conclusion : vers 7 h le matin et 16 h l'après-midi (plus ou moins 1 heure), lorsque les gens se déplacent de leur domicile vers leur entreprise, le trafic est le plus dense. De 10 h à 14 h, pendant les jours ouvrables, le volume de trafic se maintient autour de 5 000 véhicules, tandis que le week-end, il est d'environ 4 000 véhicules. Le trafic est donc similaire pour les deux types de jours pendant cette période.

WEATHER INDICATOR



Ajouter desLa conclusion provisoire est que la température n'est pas un indicateur fiable, car aucun signe ne montre une relation claire entre la température et le volume de trafic. Lorsque la température augmente, le volume de trafic peut évoluer dans les deux sens.



Il y a trois types de météo que nous devrions considérer : «shower snow», «light snow» et «proximity thunderstorm with drizzle» ; ils ont dépassé 5 000 véhicules, ce qui montre que certains types de météo peuvent avoir un impact significatif sur le trafic.

MODEL DEVELOPMENT

Division du dataset (train-test)

- Le train (80 %) est utilisé pour apprendre à l'algorithme d'apprentissage automatique à faire des prédictions précises.
- Le test (20 %) est utilisé pour évaluer la performance de l'algorithme après l'apprentissage.

METRIQUES

L'Évaluation de la performance des modèles fera en utilisant des métriques adaptées aux séries temporelles:

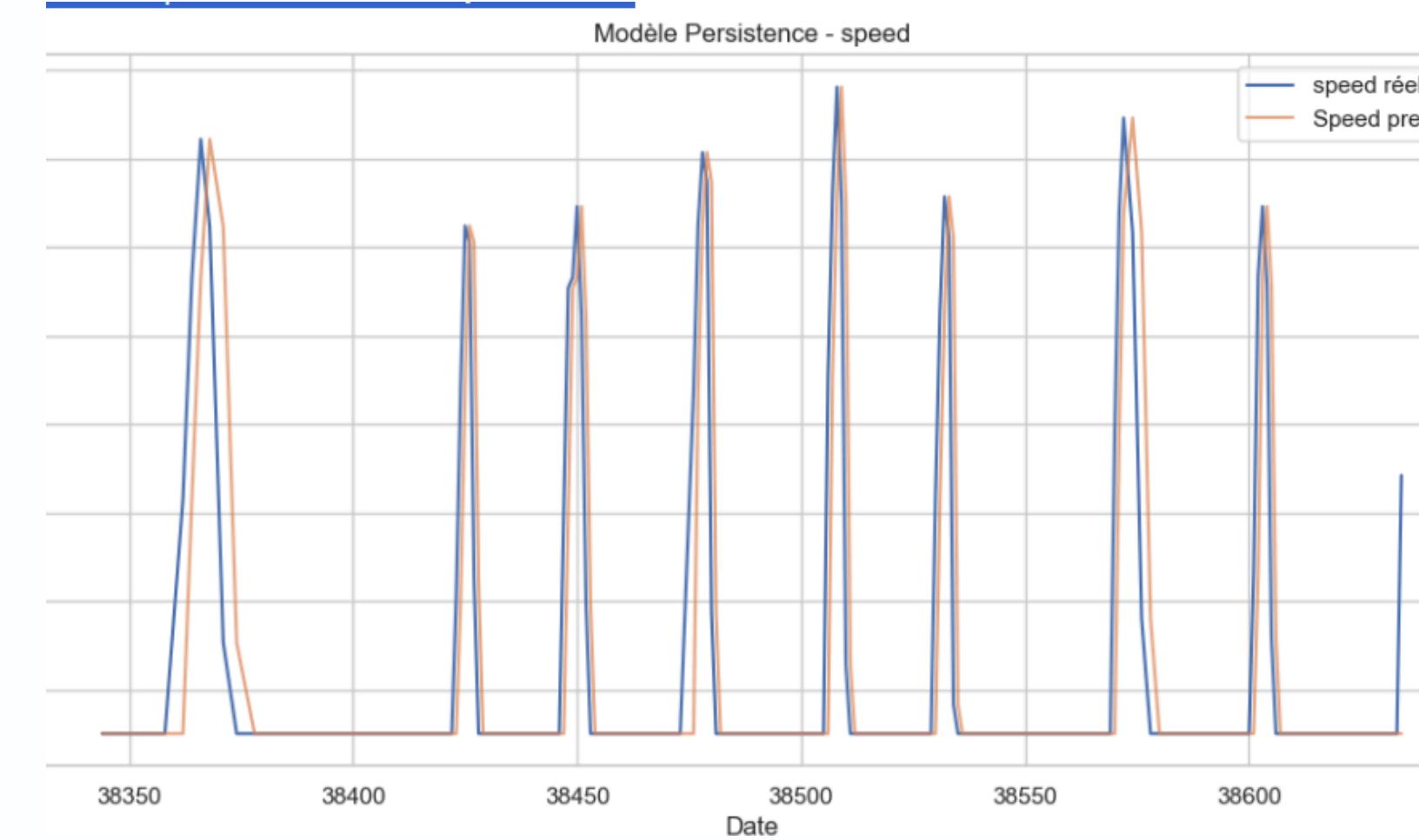
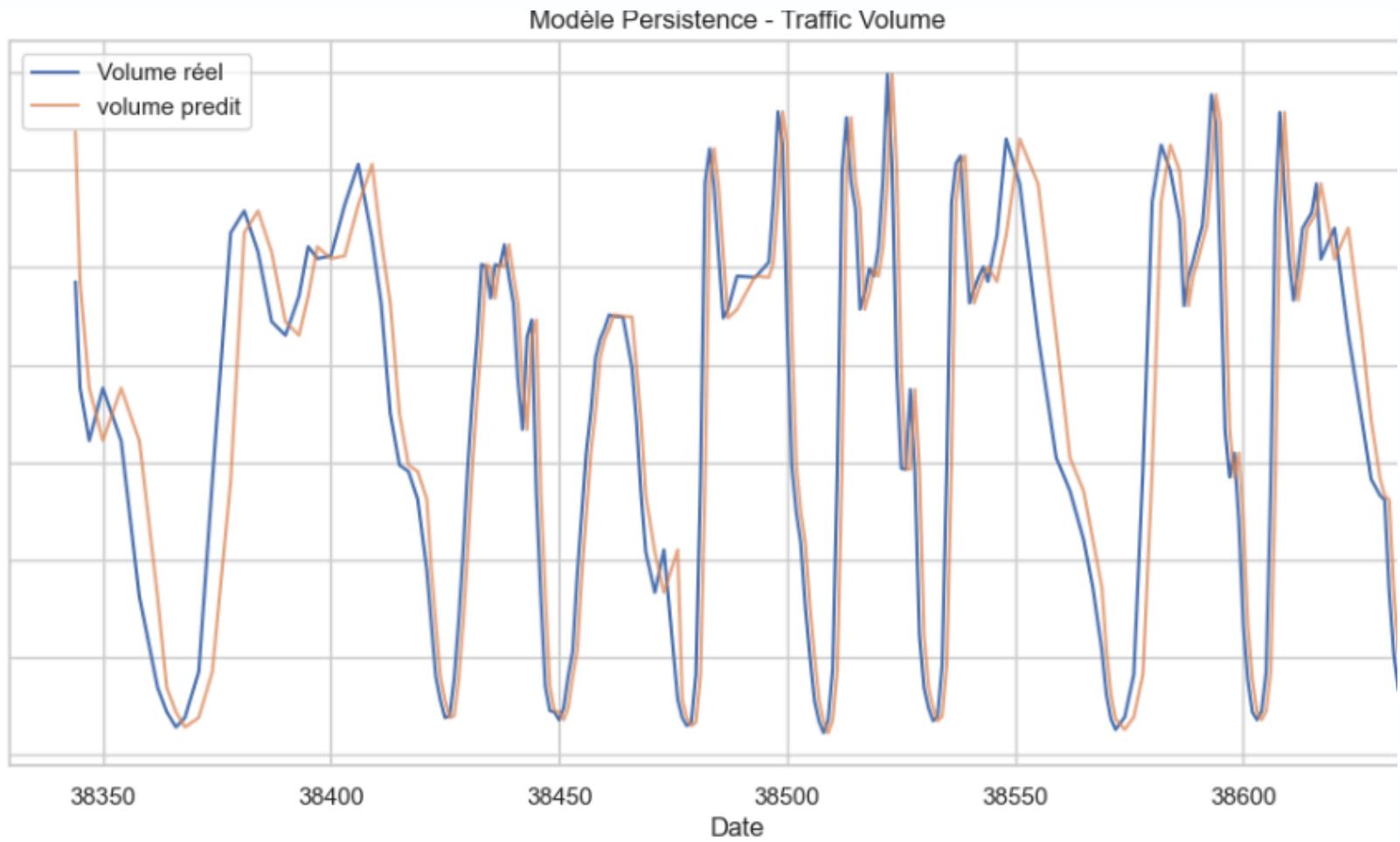
RMSE (Root Mean Squared Error)
MAE (Mean Absolute Error)

TARGET: Traffic_volume & Speed

ML MODEL

PERSISTENCE MODEL

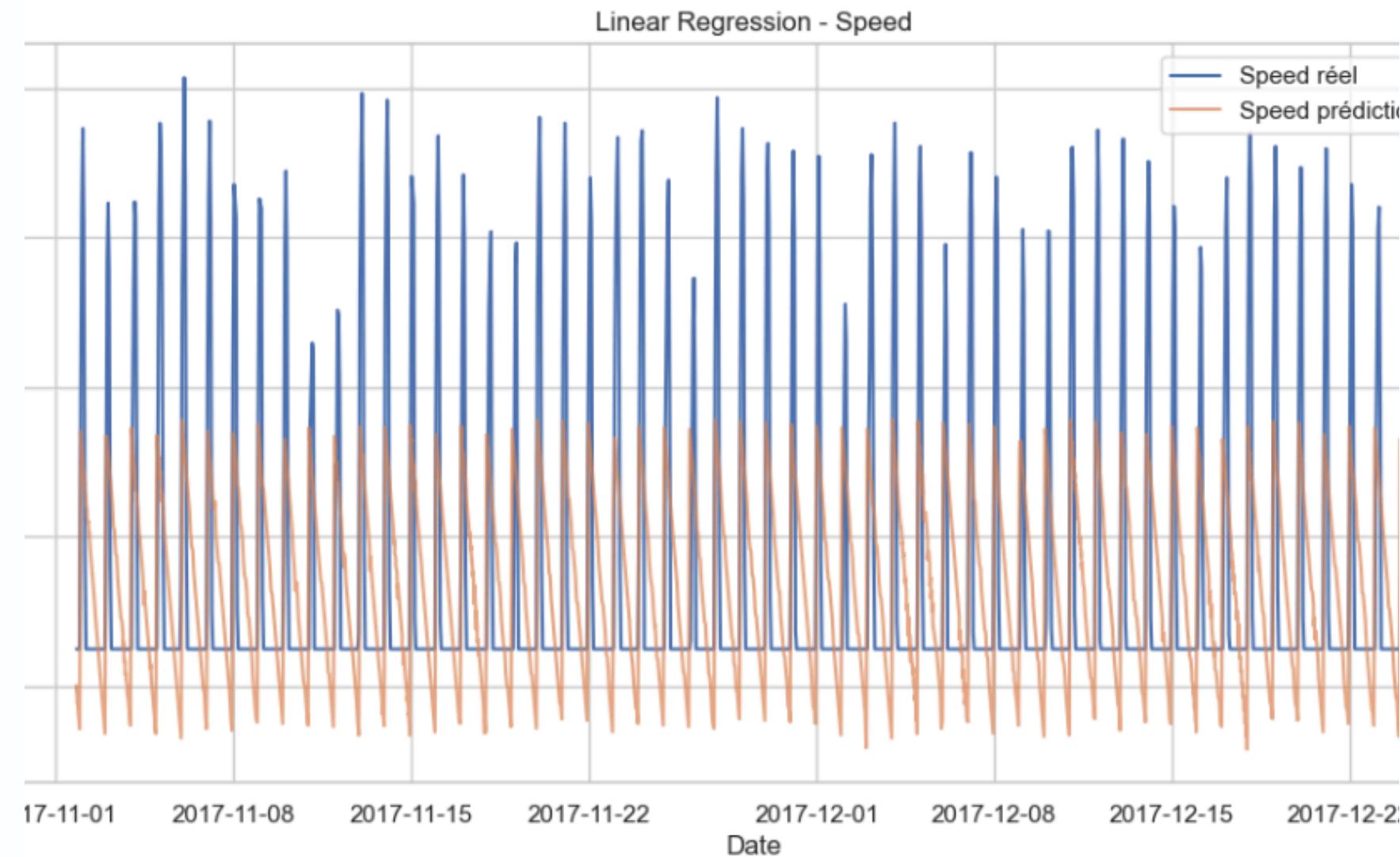
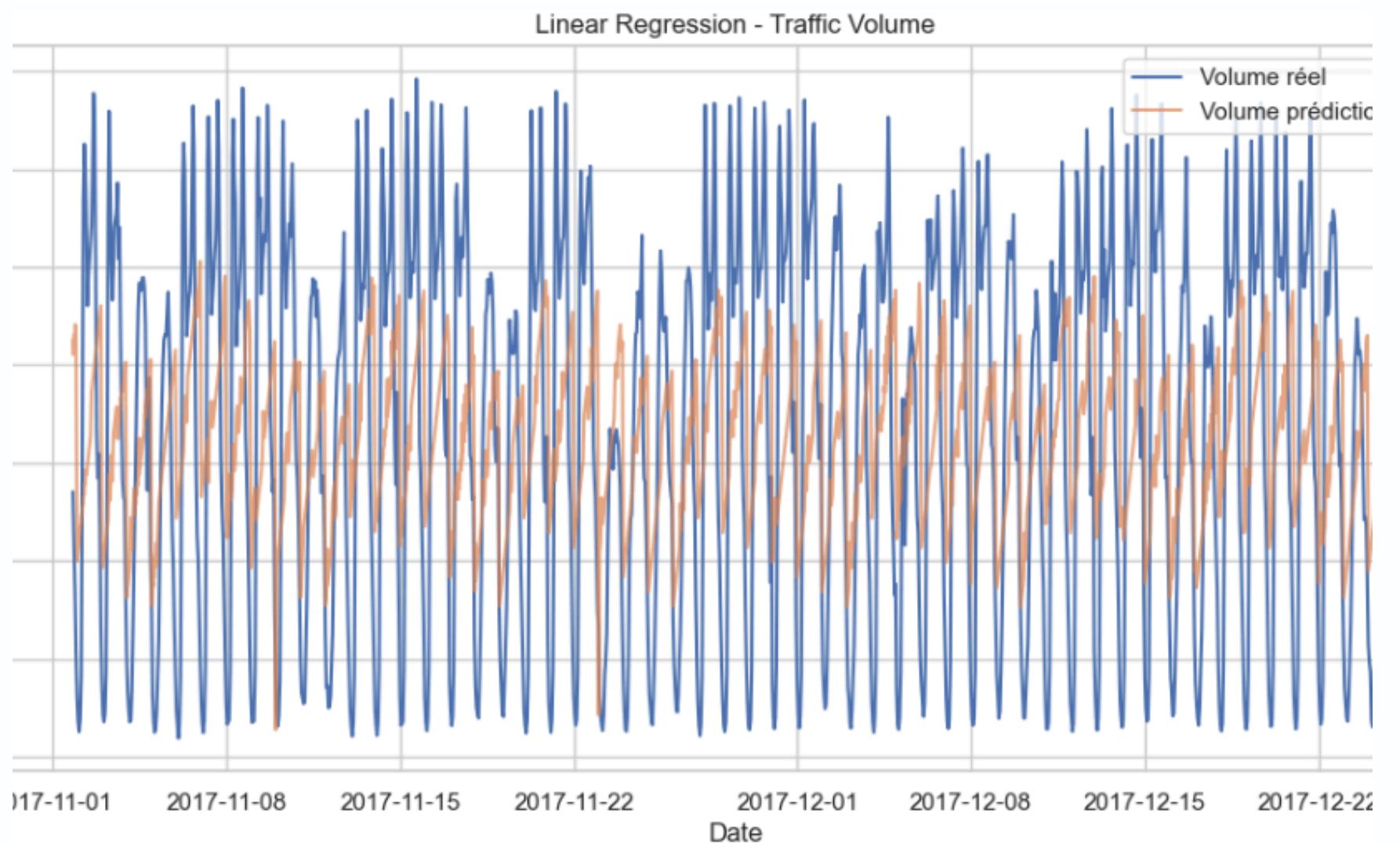
- Persistence Volume - RMSE: 811.81, MAE: 585.56
- Persistence Speed - RMSE: 13.14, MAE: 5.31



MODEL SERVANT DE BENCHMARK

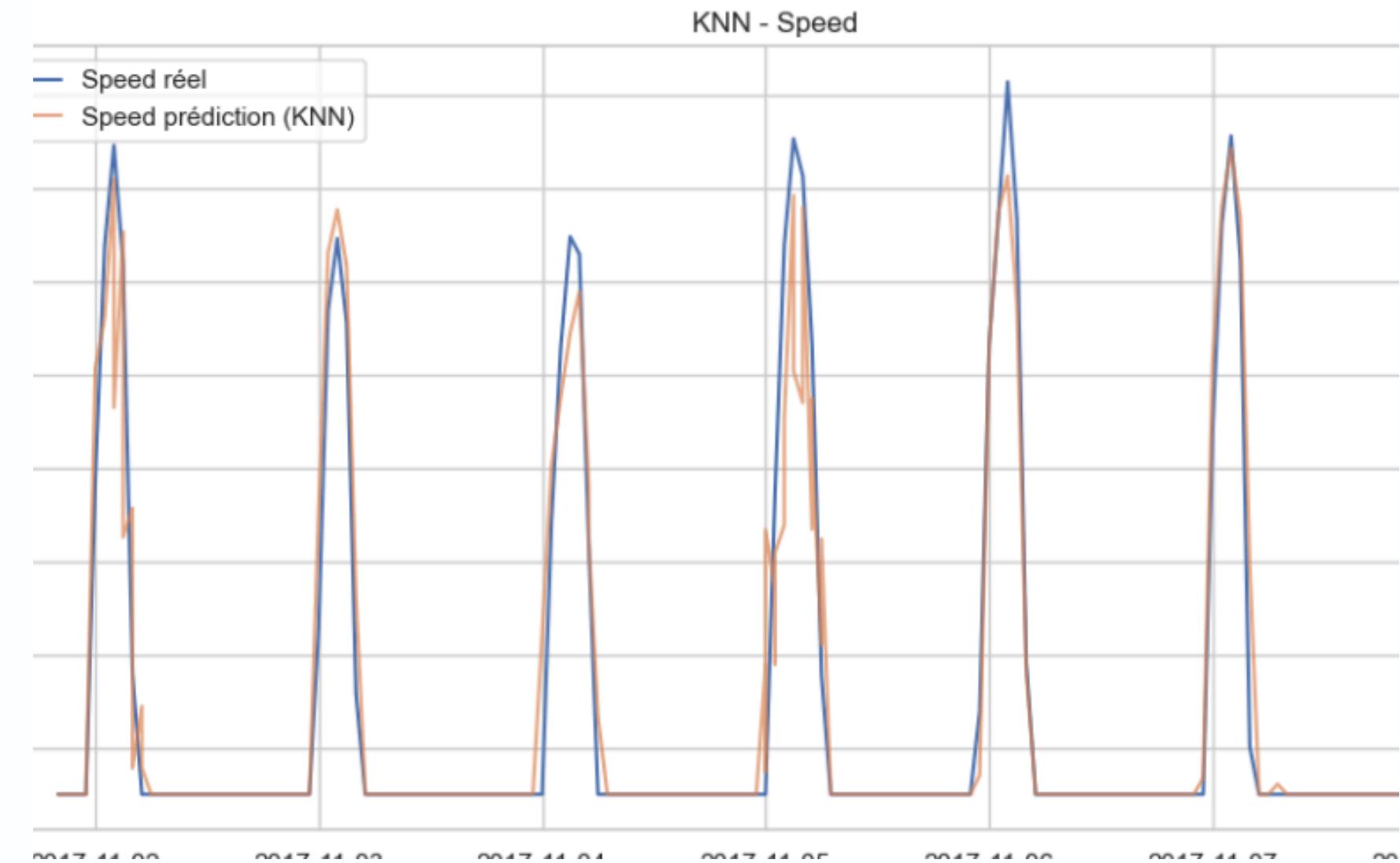
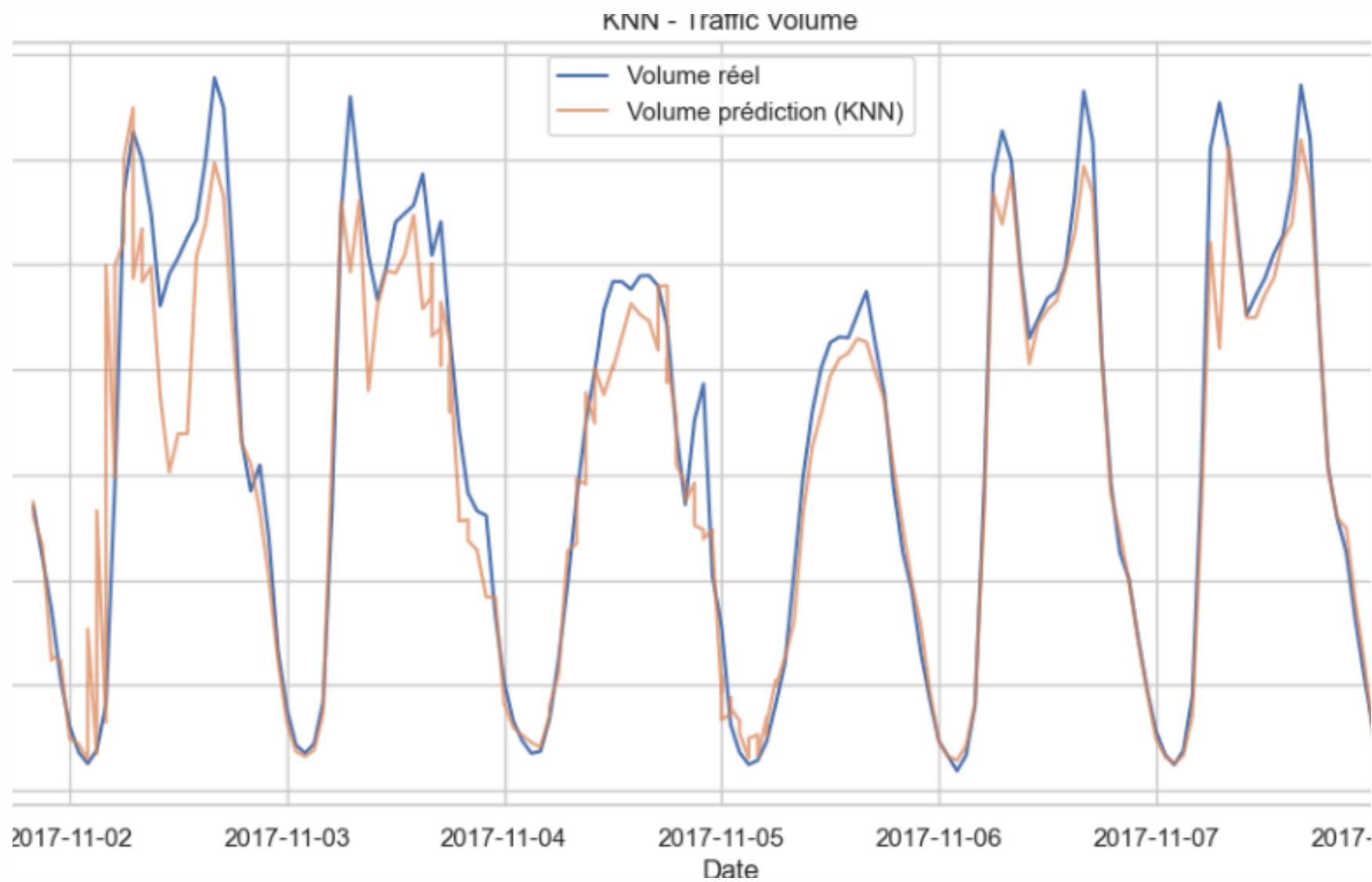
LINEAR REGRESSION

- Linear Regression Volume - RMSE: 1804.26, MAE: 1587.67
- Linear Regression Speed - RMSE: 16.23, MAE: 12.50



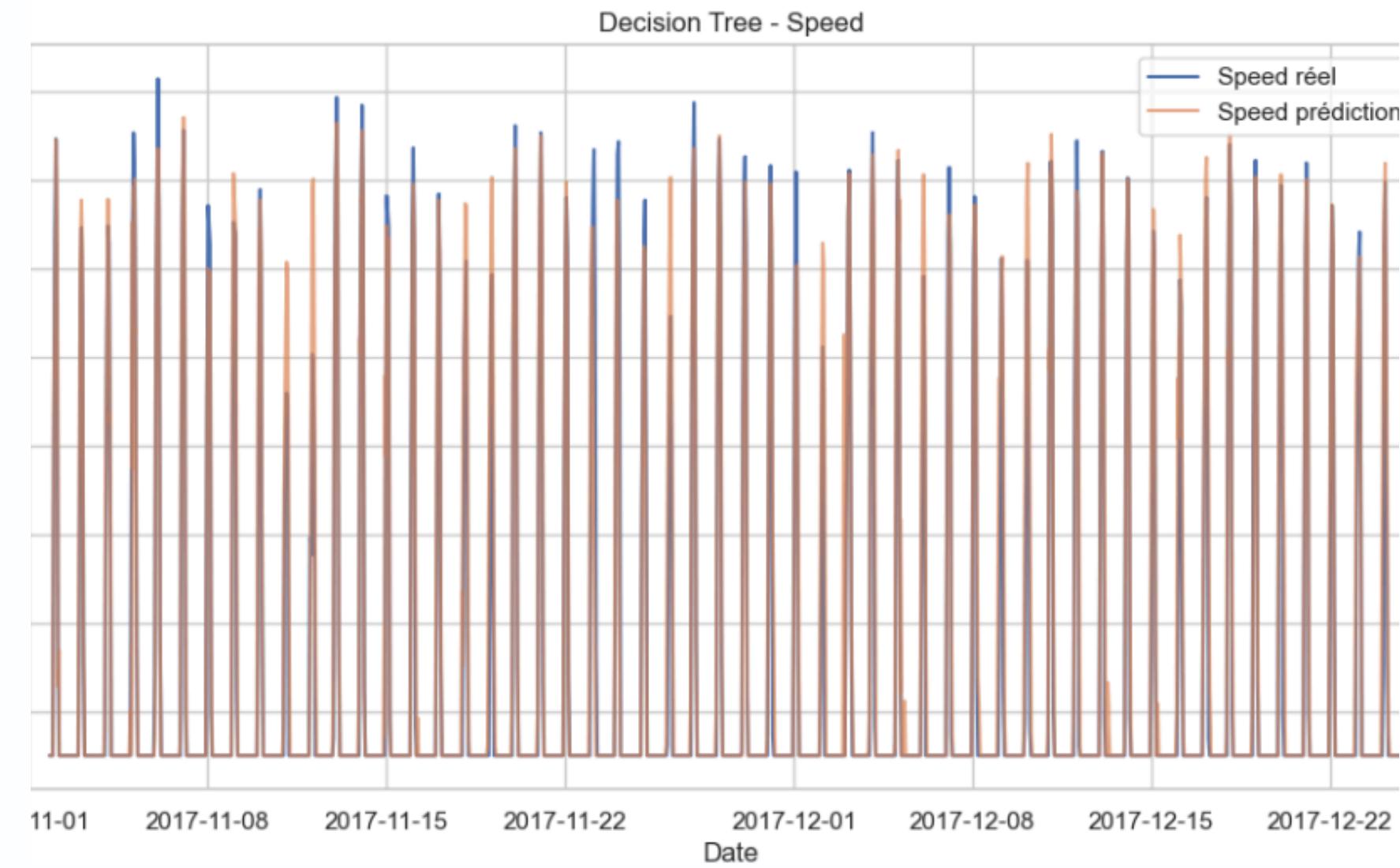
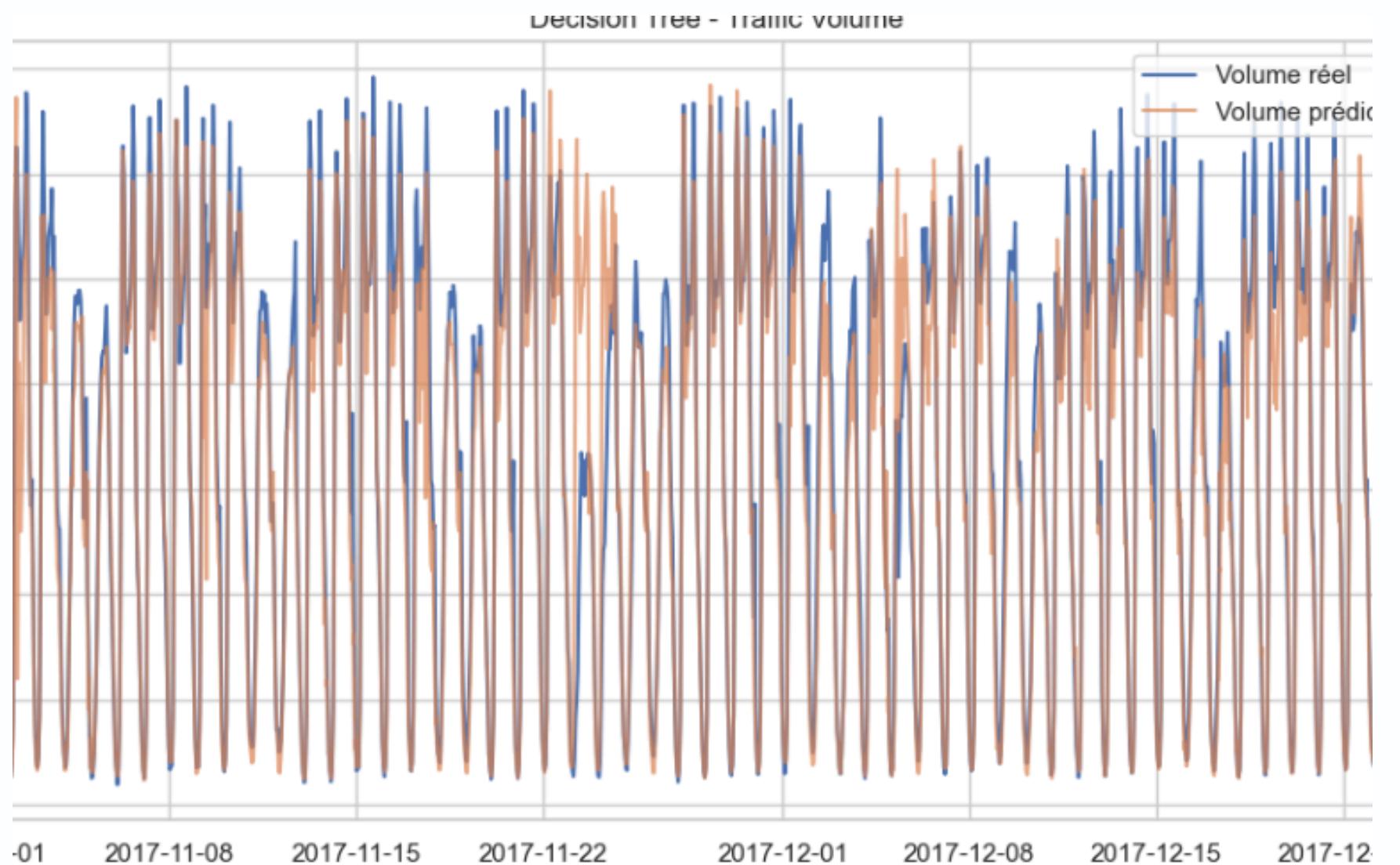
KNN :K-Nearest Neighbors

- KNN Volume - RMSE: 633.63, MAE: 387.38
- KNN Speed - RMSE: 6.95, MAE: 2.59



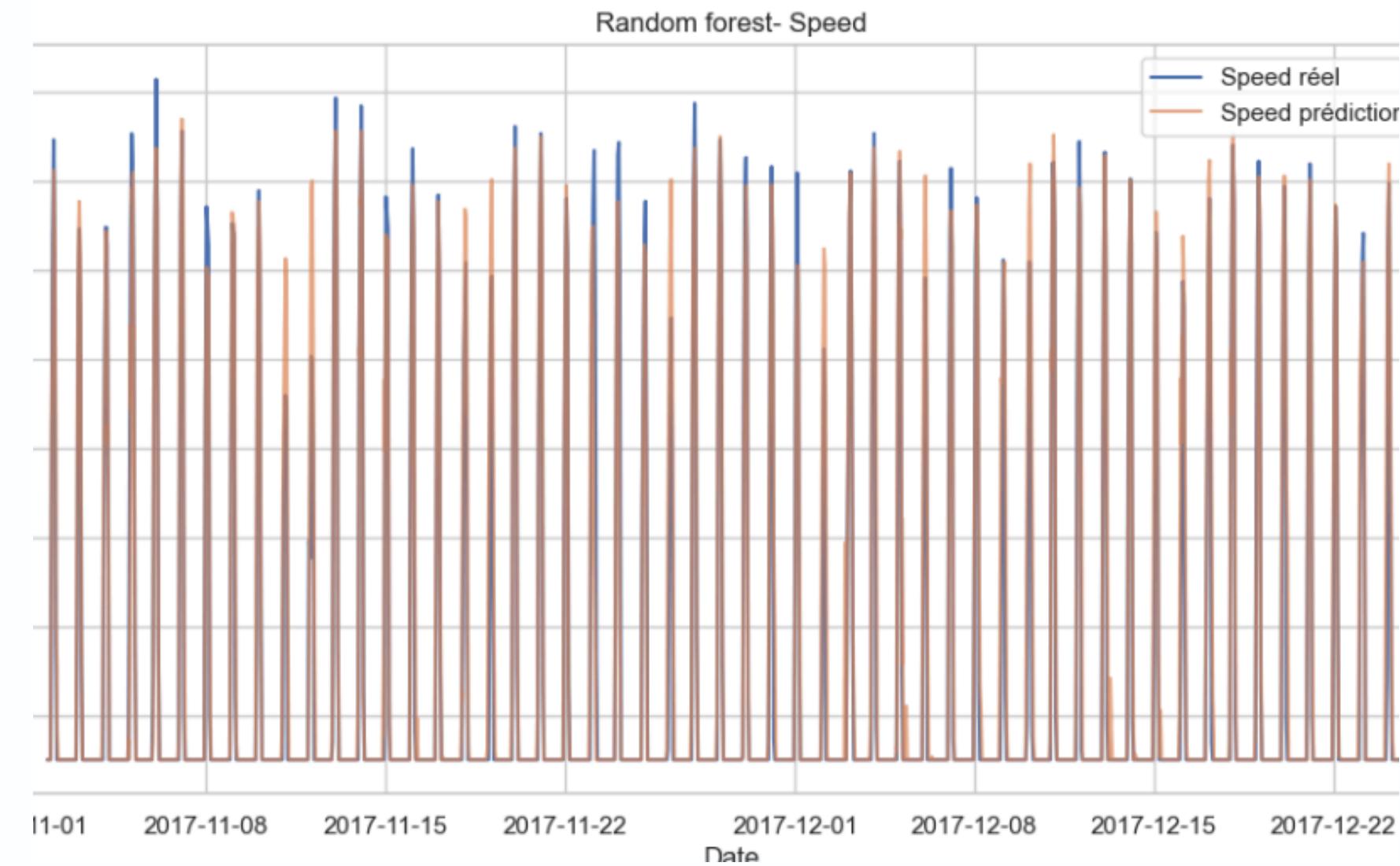
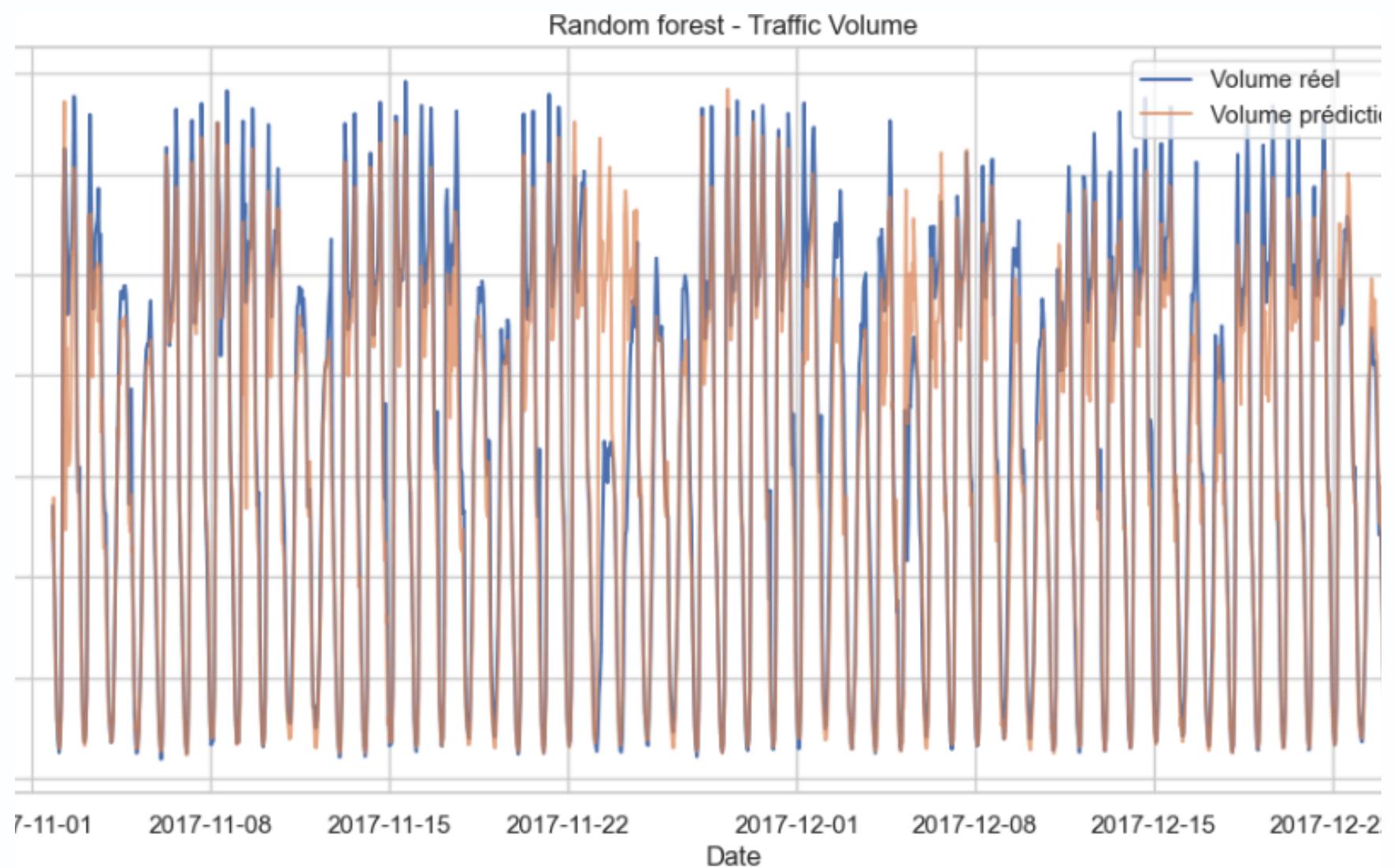
DECISION TREE

- Decision Tree Volume - RMSE: 583.58, MAE: 327.45
- Decision Tree Speed - RMSE: 5.85, MAE: 1.80



RANDOM FOREST

- Random Forest Volume - RMSE: 564.19, MAE: 316.65
- Random Forest Speed - RMSE: 5.36, MAE: 1.74



PERFORMANCE

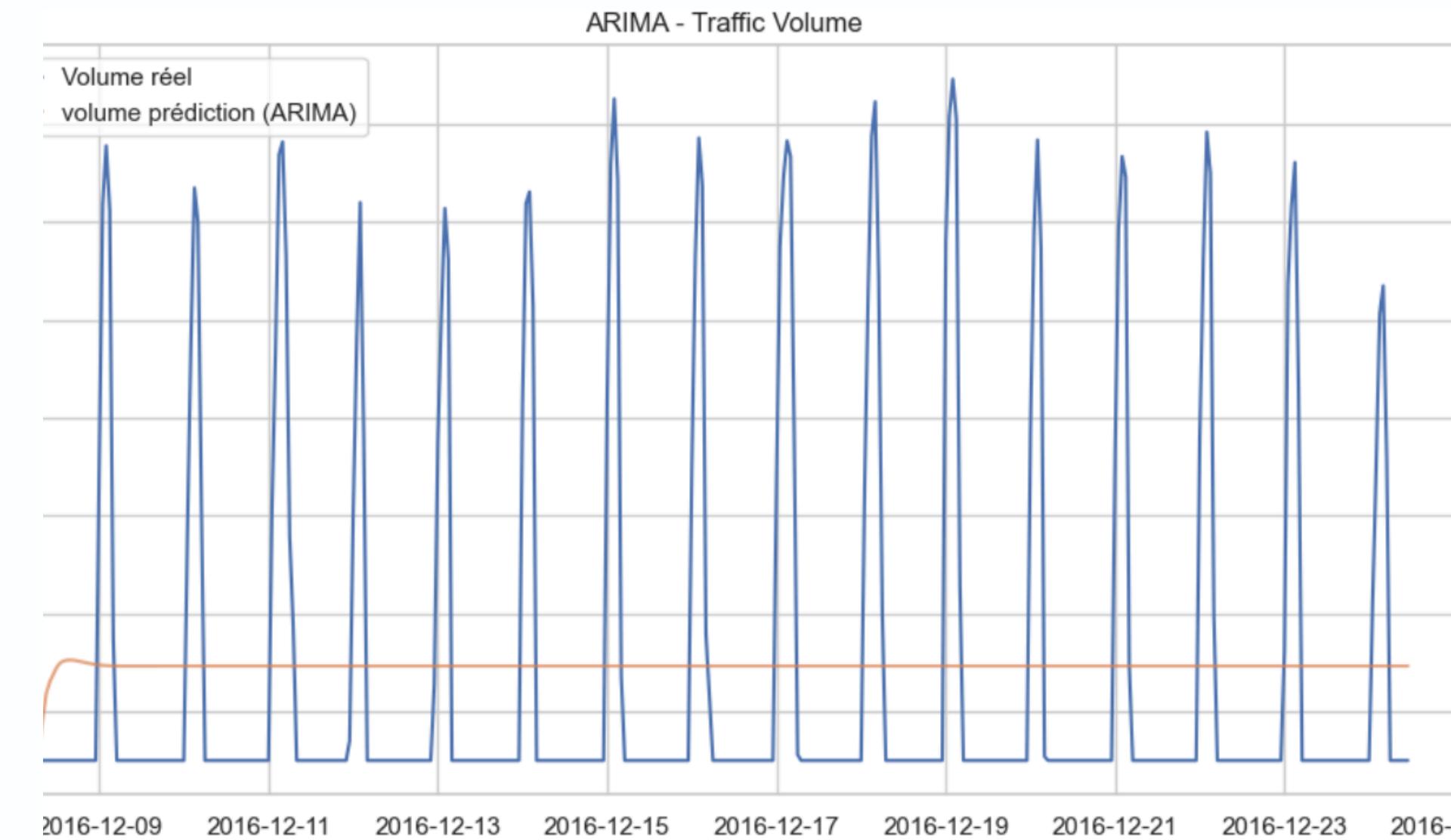
Les résultats montrent que le Random Forest est le modèle qui performe le mieux, avec des erreurs (RMSE et MAE) plus faibles que le modèle Persistence. Cette supériorité s'explique par sa capacité à exploiter plusieurs variables explicatives (heure, jour, météo, jours fériés) et à capturer des relations non linéaires entre ces variables et les cibles.

Ainsi, parmi les modèles de machine learning testés, le Random Forest constitue la meilleure approche, surpassant largement le modèle baseline (Persistence), et peut être utilisé comme référence pour toute comparaison ultérieure avec d'autres modèles plus avancés.

TIME SERIE MODEL

ARIMA

- ARIMA Volume - RMSE: 1976.11, MAE: 1733.70
- ARIMA Speed - RMSE: 19.88, MAE: 14.72



ARIMA est un modèle purement basé sur la dépendance temporelle de la série, sans prendre en compte les features exogènes comme l'heure, le jour de la semaine, les jours fériés ou la météo.

Hors le trafic routier dépend énormément de ces facteurs externes.

ARIMA ne peut «apprendre» que la tendance et la saisonnalité simple (comme un cycle quotidien ou hebdomadaire). Les variations ponctuelles (pluie, vacances, événements spéciaux) ne sont pas captées → prédictions souvent loin de la réalité → RMSE et MAE él

SARIMAX

- RMSE: 1766.0120127106597 MAE:
1559.0256422239886

Le modèle SARIMAX (Seasonal ARIMA with eXogenous regressors) est une extension de ARIMA qui permet: De modéliser la saisonnalité (SARIMA), D'inclure des variables explicatives externes (exogenous variables), appelées X.

Dans notre cas, nous avons intégré la vitesse moyenne des véhicules (speed) comme variable explicative. Cette variable est fortement corrélée avec le volume de trafic : quand le trafic augmente, la vitesse diminue, et inversement.

L'utilisation de SARIMAX avec la vitesse comme variable explicative montre l'intérêt des modèles multivariés pour la prévision du trafic. Elle illustre comment l'intégration d'informations externes pertinentes peut rendre les prédictions plus fiables et plus précises. Mais sa performance est toujours mauvaise par rapport au Random forest

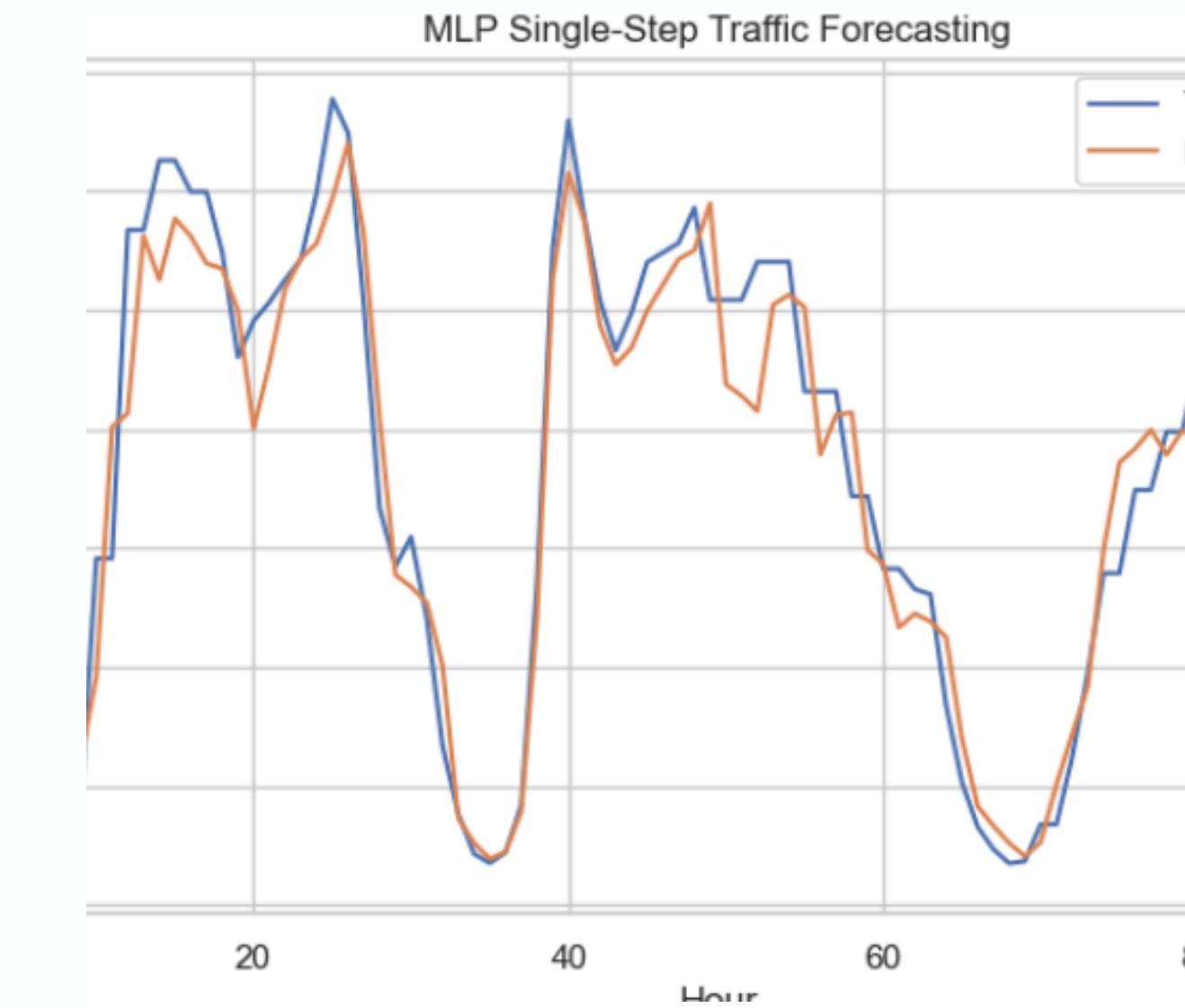
DEEP LEARNING MODEL

MLP (MULTILAYER PERCEPTRON)

- MLP Single-Step - RMSE: 453.91, MAE: 329.28

Layer (type)	Output Shape	Param #
dense_25 (Dense)	(None, 128)	9,344
dropout_14 (Dropout)	(None, 128)	0
dense_26 (Dense)	(None, 64)	8,256
dense_27 (Dense)	(None, 1)	65

MLP VA SERVIR DE BASELINE



TRANSFORMER

- Transformer Single-Step - RMSE: 442.72, MAE: 316.81



Layer (type)	Output Shape	Param #	Connected to
input_layer_14 (InputLayer)	(None, 24, 3)	0	-
multi_head_attention_1 (MultiHeadAttention)	(None, 24, 3)	963	input_layer_14[0], input_layer_14[0]
dropout_16 (Dropout)	(None, 24, 3)	0	multi_head_attention_1[0]
add_1 (Add)	(None, 24, 3)	0	dropout_16[0][0], input_layer_14[0]
layer_normalization_1 (LayerNormalization)	(None, 24, 3)	6	add_1[0][0]
flatten_3 (Flatten)	(None, 72)	0	layer_normalization_1[0]
dense_28 (Dense)	(None, 64)	4,672	flatten_3[0][0]
dropout_17 (Dropout)	(None, 64)	0	dense_28[0][0]
dense_29 (Dense)	(None, 1)	65	dropout_17[0][0]

CNN

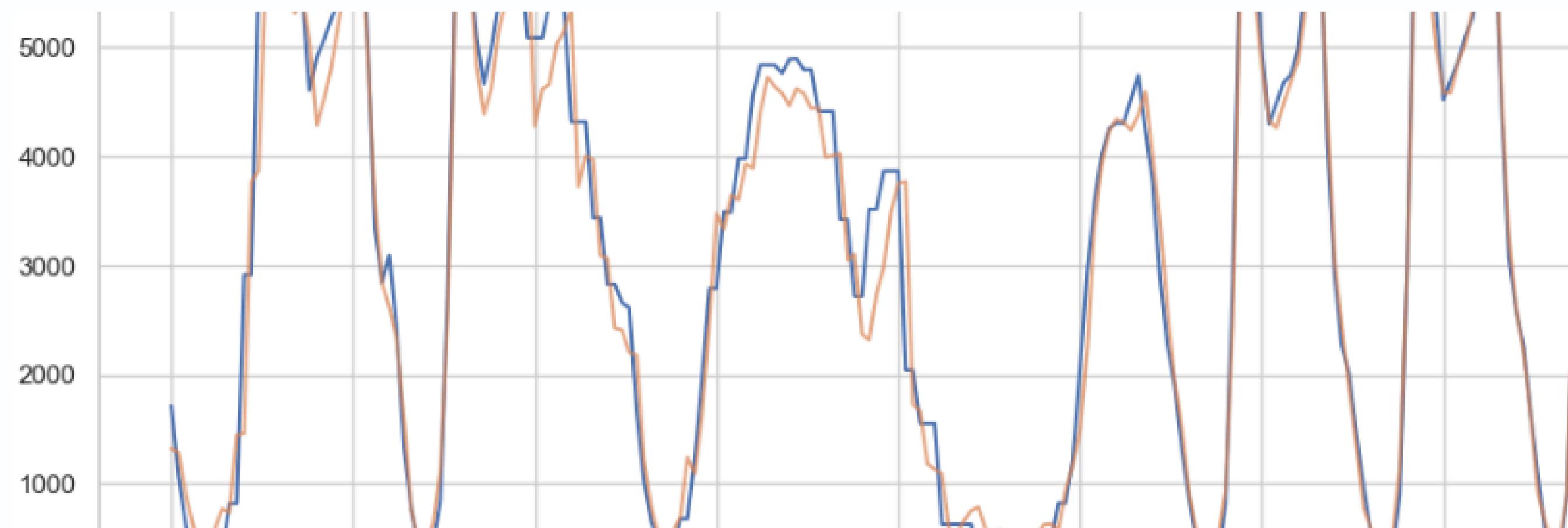
- CNN Single-step - RMSE: 440.04, MAE: 318.57

Layer (type)	Output Shape	Param #
conv1d_4 (Conv1D)	(None, 22, 64)	640
max_pooling1d_4 (MaxPooling1D)	(None, 11, 64)	0
flatten_4 (Flatten)	(None, 704)	0
dropout_18 (Dropout)	(None, 704)	0
dense_30 (Dense)	(None, 32)	22,560
dense_31 (Dense)	(None, 1)	33

LSTM

- LSTM Single-step - RMSE: 417.46, MAE: 296.62

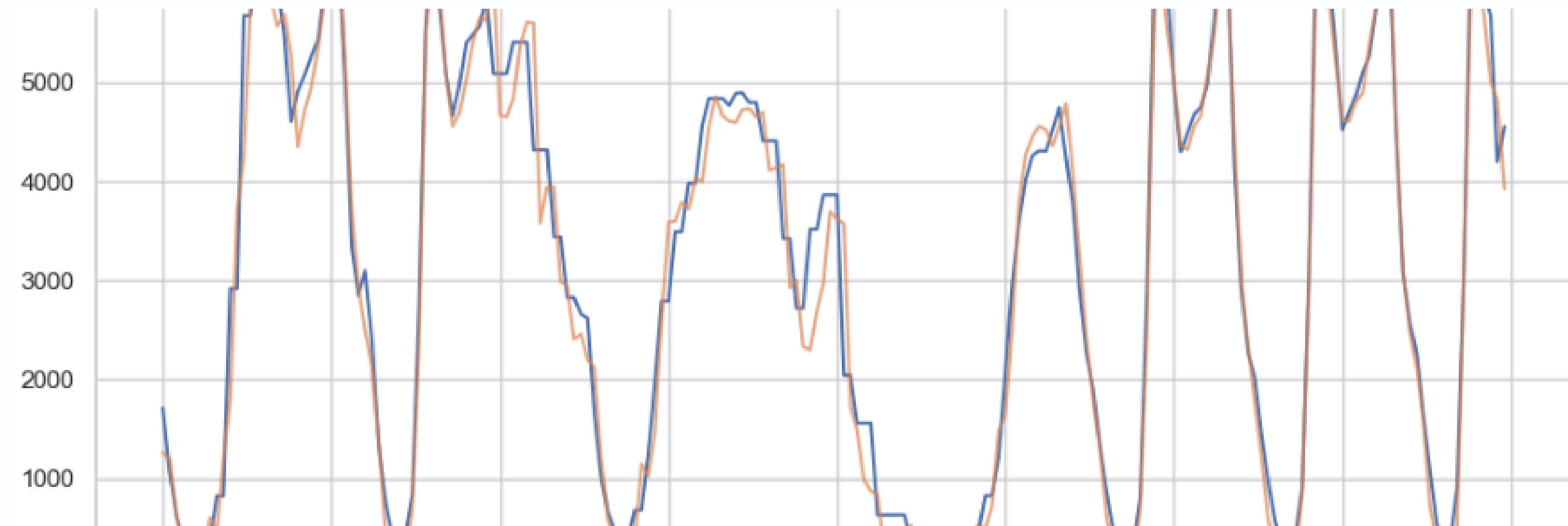
Layer (type)	Output Shape	Param #
lstm_5 (LSTM)	(None, 64)	17,408
dropout_19 (Dropout)	(None, 64)	0
dense_32 (Dense)	(None, 32)	2,080
dense_33 (Dense)	(None, 1)	33



GRU (Gated Recurrent Unit)

- GRU Single-step - RMSE: 407.64, MAE: 288.96

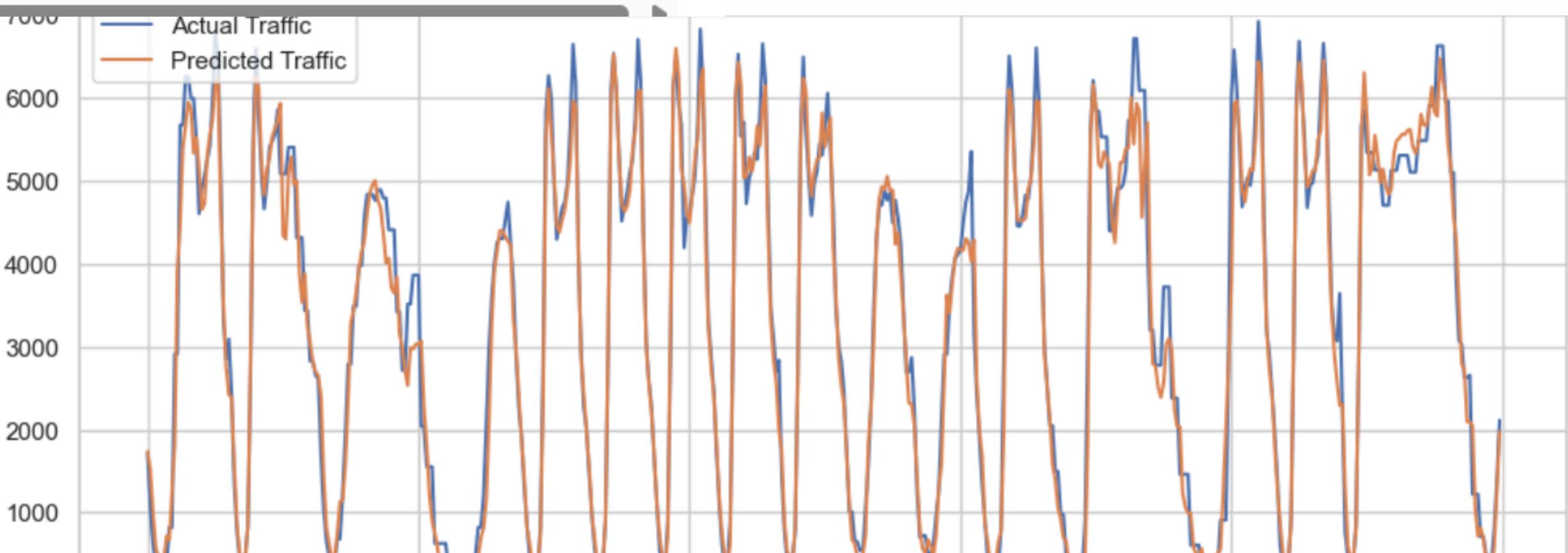
Layer (type)	Output Shape	Param #
gru_3 (GRU)	(None, 64)	13,248
dropout_20 (Dropout)	(None, 64)	0
dense_34 (Dense)	(None, 32)	2,080
dense_35 (Dense)	(None, 1)	33



CNN-LSTM

- RMSE = 376.91897545767193

Layer (type)	Output Shape	Param #
conv1d_8 (Conv1D)	(None, 22, 64)	1,216
max_pooling1d_8 (MaxPooling1D)	(None, 11, 64)	0
lstm_10 (LSTM)	(None, 64)	33,024
dropout_28 (Dropout)	(None, 64)	0
dense_50 (Dense)	(None, 32)	2,080
dense_51 (Dense)	(None, 1)	33



MULTISTEP

ON VA PREDIRE LES 6 PROCHAINES HEURES

Horizon +1h - RMSE: 426.72, MAE: 310.06

Horizon +2h - RMSE: 566.85, MAE: 400.06

Horizon +3h - RMSE: 688.50, MAE: 473.63

Horizon +4h - RMSE: 779.99, MAE: 534.31

Horizon +5h - RMSE: 859.55, MAE: 598.99

Horizon +6h - RMSE: 940.51, MAE: 663.48

Multi-step CNN-LSTM - RMSE moyen: 710.35, MAE moyen: 496.76

RESULT & CONCLUSION

RESULT

- Les modèles ML classiques (Random Forest, Decision Tree, KNN) fournissent une base de comparaison.
- Random Forest est le meilleur parmi eux :
- Volume de trafic : RMSE ≈ 564 , MAE ≈ 317
- Vitesse : RMSE ≈ 5.36 , MAE ≈ 1.74
- Les modèles de séries temporelles (ARIMA, SARIMAX) tiennent compte de la dépendance temporelle mais sont moins performants que Random Forest.
- Les modèles Deep Learning, en particulier CNN-LSTM, offrent la meilleure performance globale :
- Prédiction single-step : RMSE Volume ≈ 389
- Prédiction multi-step (6h) : RMSE Volume ≈ 697 , MAE ≈ 486

Conclusion

- L'intégration de caractéristiques temporelles et exogènes (vitesse, météo, etc.) améliore considérablement les prédictions.
- CNN-LSTM est le modèle le plus performant pour la prévision du trafic.
- Ce projet démontre le potentiel du Deep Learning pour les systèmes de transport intelligents (ITS) et la gestion du trafic en temps réel.



PERSPECTIVES

- Ajouter des features supplémentaires (accidents, événements, feux de circulation) pour des prédictions plus précises.
- Déployer une interface web pour la prévision du trafic en temps réel.



THANK YOU