MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE RECHERCHE

REPUBLIQUE DE COTE D'IVOIRE

Institut Supérieur de Statistique D'Econométrie

Union-Discipline-Travail

MASTER 1

STATISTIQUES – ECONOMETRIE – DATA SCIENCE

MINI PROJET

ANALYSE STATISTIQUES ECONOMETRIQUES

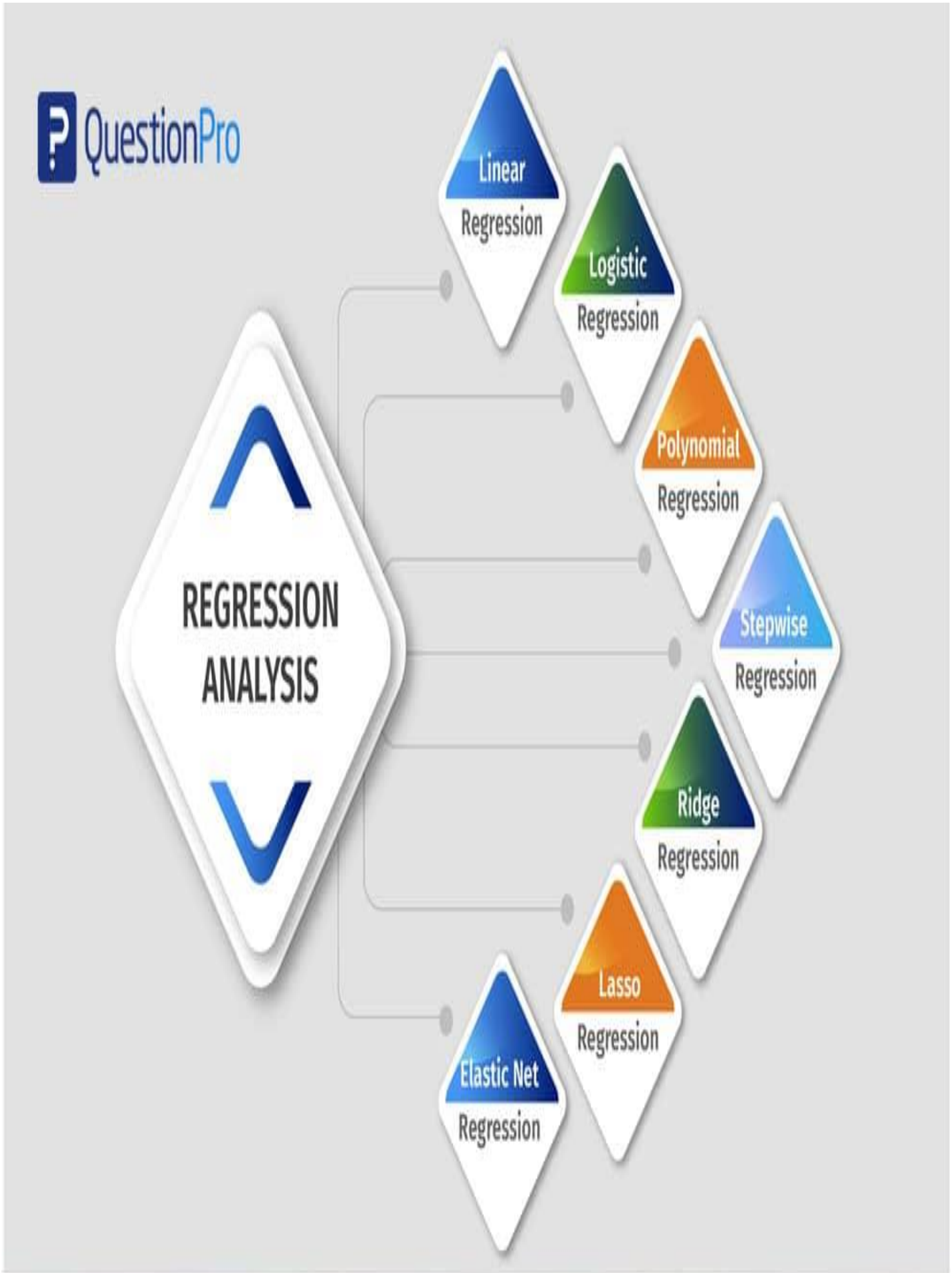
MODÉLISATION DES DÉLAIS DE LIVRAISON DE NOURRITURE : APPROCHE PAR RÉGRESSION LINÉAIRE MULTIPLE ET ANALYSE DE LA VARIANCE (ANOVA)

Nom: YOBO

Prénom(s): BAYE GUY ANGE HENOC

Enseignant – Encadreur

AKPOSSO DIDIER MARTIAL


QuestionPro
REGRESSION ANALYSIS
Linear Regression
Logistic Regression
Polynomial Regression
Stepwise Regression
Ridge Regression
Lasso Regression
Elastic Net Regression

AVANT-PROPOS

Ce projet vise à analyser et optimiser les délais de livraison dans le secteur de la livraison de nourriture. En étudiant les facteurs influençant ces délais, tels que la distance, la météo, les conditions de circulation, l'heure de la journée, le type de véhicule, l'expérience du coursier et le temps de préparation des commandes, il cherche à prédire le temps total de livraison. À travers l'exploration du dataset ***delai_livraison.csv***, l'objectif est de développer un modèle de prédiction basé sur des méthodes d'apprentissage automatique. Ce projet offre une occasion unique de mieux comprendre la logistique des livraisons et de proposer des solutions pratiques pour améliorer l'efficacité des entreprises de livraison.

Table des matières

AVANT-PROPOS ... 3
INTRODUCTION ... 7
PARTIE I : ANALYSE DESCRIPTIVE ... 8
I) APPROCHE METHODOLOGIQUE DES DONNEES ... 8
1) Présentation du dictionnaire de données ... 8
2) Présentation du jeu de données ... 9
3) Détection et apurement des valeurs Manquantes et Aberrantes / extrême 10
3.1) Visualisation et traitements des valeurs manquantes ... 10
3.2) Visualisation et traitements des valeurs aberrantes / et ou Extrêmes ... 11
II) ANALYSE EXPLORATOIRE DES DONNEES (EDA) ... 12
1) Analyses Univariées ... 12
1.1) Paramètres statistiques ... 12
1.2) Interprétation des paramètres du tableau statistique ... 13
1.3) Graphiques des variables numériques ... 15
1.4) Interprétation des graphiques des variables numeriques ... 15
2) Analyse Bivariees ... 17
2.1) Matrice de corrélation ... 17
PARTIE II : REGRESSION LINEAIRE MULTIPLES ... 19
1) Construction du modèle de régression linéaire multiple ... 20
1.1) Partitionnement des données en ensemble d'entraînement et de test ... 20
1.2) Représentation des variables ... 20
1.3) Modèle de régressions linéaires ... 21
1.4) Estimation des paramètres du modèle ... 21
1.5. Calcul des métriques de performance du modèle ... 25
2) Analyse Résiduels et Validation du Modèle de Régression ... 26
2.1) Analyse des résidus ... 26
2.2) Validation du Modèle de Régression ... 27
3) Prédiction d'une nouvelle valeur ... 30
PARTIE III : ANOVA ... 31
1) Graphe des délais de livraison selon la météo ... 31
2) Estimation des statistiques de base (mean, quantile, sd) ... 31
3) Teste de la normalité des données dans chaque temps météorologique. ... 32

3.1) Normalité 32

4) Teste de l'égalité des variances 33

5) Faire un test robuste par bootsrap (rééchantillonage) 34

6) Tester la significativité du facteur : tester l'égalité des moyennes 34

7) Analyser les résidus 35

8) Interpréter les coefficients 35

9) Modèle ANOVA 35

Tableau de bord 36

CONCLUSION 37

ANNEXE 38

Liste des tableaux

Tableau 1 Dictionnaire des données 8
Tableau 2 Extraire du jeu de donnée 9
Tableau 3 Information supplémentaires sur les variables 9
Tableau 4 Paramètre statistiques 13
Tableau 5 Partitionnement 20
Tableau 6 Estimation statistiques 31
Tableau 7 Shapiro test 32
Tableau 8 Boostrap 34

Liste des figures

Figure 1 visualisations des valeurs manquantes 10
Figure 2 Traitements des valeurs manquantes (imputation) 11
Figure 3 Visualisation des valeurs aberrantes 11
Figure 4 Traitement des valeurs manquantes 12
Figure 5 Variables numériques 15
Figure 6 Matrice de corrélation 17
Figure 7 Représentation des variables 20
Figure 8 Modèle de régression 21
Figure 9 Estimation des coefficients 22
Figure 10intervalle de confiance 22
Figure 11 Results_estimation 23
Figure 12choix des variables 24
Figure 13 Métriques de performances 25
Figure 14 Résidus 26
Figure 15 Linéarité 27
Figure 16 Multi colinéarité 27
Figure 17 Auto corrélation d'erreur 28
Figure 18 RESIDUS VS VALEURS PREDITES 29
Figure 19 Analyse prévisionnelle 30
Figure 20 Délais de livraison en fonction de la météo 31
Figure 21 représentation des temps 32
Figure 22Analyse des résidus 35

INTRODUCTION

Contexte et justification de l'étude

Le secteur de la livraison de nourriture connaît une croissance rapide, soutenue par la demande croissante des consommateurs pour des services rapides et efficaces. Cependant, l'un des plus grands défis auquel les entreprises de livraison font face est la gestion des délais de livraison, qui peut être affectée par une multitude de facteurs tels que la distance, la météo, les conditions de circulation et l'heure de la journée. Optimiser ces délais est non seulement crucial pour améliorer l'expérience client, mais aussi pour réduire les coûts opérationnels et maximiser l'efficacité des courriers. Dans ce contexte, cette étude se propose d'analyser un ensemble de données détaillant les différents paramètres influençant les délais de livraison, afin de développer un modèle de prédiction fiable et applicable dans le secteur.

Problématique

La question centrale de cette étude réside dans l'identification des facteurs clés qui influencent le délai de livraison et la capacité à prédire ces délais avec une précision suffisante pour améliorer l'efficacité des services de livraison. Plus précisément, il s'agit de comprendre comment des éléments comme la distance, la météo, le trafic, le type de véhicule et l'expérience du coursier peuvent interagir pour affecter le temps de livraison global. L'enjeu est de déterminer si ces facteurs peuvent être modélisés de manière à prédire le délai de livraison de manière efficace et fiable, en vue d'optimiser les processus logistiques et réduire les coûts.

Principaux résultats attendus

Les principaux résultats attendus de cette étude sont les suivants :

1. Une meilleure compréhension des facteurs ayant le plus grand impact sur les délais de livraison.
2. Le développement d'un modèle prédictif capable de calculer avec précision le temps de livraison en fonction des caractéristiques observées.
3. Des recommandations pratiques pour les entreprises de livraison visant à réduire les délais de livraison, notamment par l'optimisation des ressources et la gestion proactive des facteurs influents.

Méthodologie

La méthodologie adoptée pour cette étude s'articule autour de plusieurs étapes clés. Tout d'abord, un prétraitement des données sera effectué afin de gérer les valeurs manquantes, de normaliser les variables numériques et de transformer les variables catégorielles en formats compatibles avec les algorithmes d'apprentissage automatique. Les techniques de régression, telles que la régression linéaire multiple et les forêts aléatoires, seront ensuite utilisées pour identifier les relations entre les variables indépendantes (distance, météo, trafic, etc.) et la variable cible (délai de livraison). La sélection de ces techniques repose sur leur capacité à traiter des relations complexes et non linéaires dans les données. Enfin, les résultats seront évalués à l'aide de métriques de performance classiques, telles que l'erreur quadratique moyenne (RMSE) et le coefficient de détermination (R^2), afin d'assurer la robustesse et la fiabilité du modèle.

PARTIE I : ANALYSE DESCRIPTIVE

I) APPROCHE METHODOLOGIQUE DES DONNEES

L'approche méthodologique des données englobe l'organisation, la collecte, l'analyse et l'interprétation des données dans le cadre d'une étude ou d'une recherche. Elle repose sur un ensemble de principes, de techniques et de processus visant à traiter les données de manière systématique et rigoureuse, afin d'obtenir des résultats fiables et pertinents.

1)Présentation du dictionnaire de données

Un dictionnaire de données est un document qui décrit en détail chaque variable utilisée dans une analyse statistique ou économétrique. Il explique les propriétés, les caractéristiques et le contexte de chaque variable, en clarifiant leur signification et leur rôle dans l'analyse.

Order_ID	Qualitative	Identifiant unique pour chaque commande.	**Numérique ou alphanumérique (Exemples : "ORD12345", "ORD67890")**
Distance_km	Quantitative	La distance de livraison en kilomètres entre le point de départ et la destination.	**Numérique (Exemples : 2.5, 10.7, 15.3, 7.9)**
Météo	Qualitative	Conditions météorologiques pendant la livraison.	**"Clair", "Pluvieux", "Neigeux", "Brumeux", "Venteux"**
Traffic_Level	Qualitative	Conditions de trafic pendant la livraison.	**"Faible", "Moyenne", "Élevée"**
Time_of_Day	Qualitative	Période de la journée durant laquelle la livraison a eu lieu.	**"Matin", "Après-midi", "Soir", "Nuit"**
Vehicle_Type	Qualitative	Type de véhicule utilisé pour la livraison.	**"Vélo", "Scooter", "Voiture"**
Preparation_Time_min	Quantitative	Temps nécessaire à la préparation de la commande en minutes.	**Numérique (Exemples : 10, 20, 25, 30)**
Courier_Experience_yrs	Quantitative	Nombre d'années d'expérience du coursier.	**Numérique (Exemples : 1, 2, 5, 10)**
Delivery_Time_min (cible)	Quantitative	Délai total de livraison en minutes, variable cible.	**Numérique (Exemples : 30, 45, 60, 75)**

Tableau 1 Dictionnaire des données

2)Présentation du jeu de données

Order_ID	Distance_km	Weather	Traffic_Level	Time_of_Day	Vehicle_Type	Preparation_Time_min	Courier_Experience_yrs	Delivery_Time_min
0	7.93	Windy	Low	Afternoon	Scooter	12	1.0	43
1	16.42	Clear	Medium	Evening	Bike	20	2.0	84
2	9.52	Foggy	Low	Night	Scooter	28	1.0	59
3	7.44	Rainy	Medium	Afternoon	Scooter	5	1.0	37
---	---	---	---	---	---	---	---	---
995	8.50	Clear	High	Evening	Car	13	3.0	54
996	16.28	Rainy	Low	Morning	Scooter	8	9.0	71
997	15.62	Snowy	High	Evening	Scooter	26	2.0	81
998	14.17	Clear	Low	Afternoon	Bike	8	0.0	55
999	6.63	Foggy	Low	Night	Scooter	N/A	N/A	N/A

Tableau 2 Extraire du jeu de donnée

L'ensemble de données présente 1000 observations pour 9 variables. Il semble qu'il y ait des valeurs manquantes dans les données.

Column	Value
Order_ID	1000
Distance_km	1000
Weather	970
Traffic_Level	970
Time_of_Day	970
Vehicle_Type	1000
Preparation_Time_min	1000
Courier_Experience_yrs	970
Delivery_Time_min	1000

Column	Dtype
Order_ID	object
Distance_km	float64
Weather	object
Traffic_Level	object
Time_of_Day	object
Vehicle_Type	object
Preparation_Time_min	int64
Courier_Experience_yrs	float64
Delivery_Time_min	int64

Tableau 3 Information supplémentaires sur les variables

Dans cette partie nous observons le nombre exact d'élément dans chaque variable et le types des variables.

3) Détection et apurement des valeurs Manquantes et Aberrantes / extrême

Dans cette section, nous allons chercher à identifier visuellement les éventuelles valeurs manquantes dans notre jeu de données, puis à les traiter. Ces valeurs peuvent provenir d'erreurs de mesure, de saisie, de calcul ou bien de valeurs extrêmes réelles présentes dans les données. Les valeurs atypiques peuvent avoir un impact majeur sur les résultats des analyses statistiques, en faussant des indicateurs comme la moyenne ou l'écart-type, mais aussi en influençant les tests d'hypothèse. Il est donc crucial de détecter et de traiter ces valeurs extrêmes avant de procéder à toute analyse statistique.

3.1) Visualisation et traitements des valeurs manquantes

a) Visualisation des valeurs manquantes

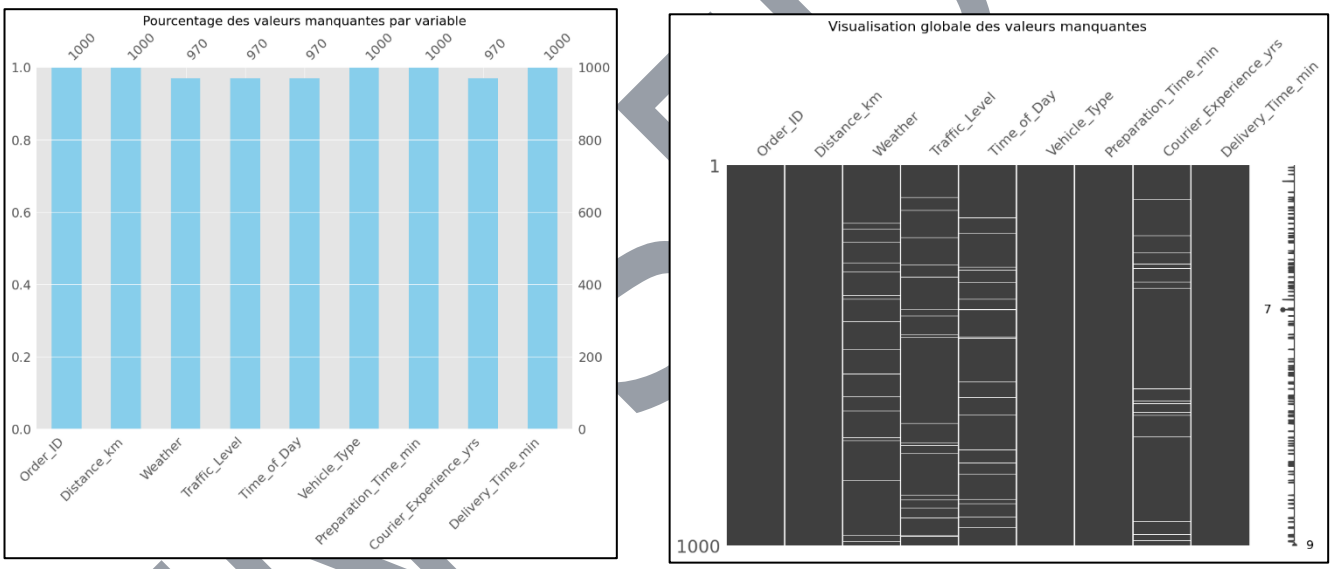
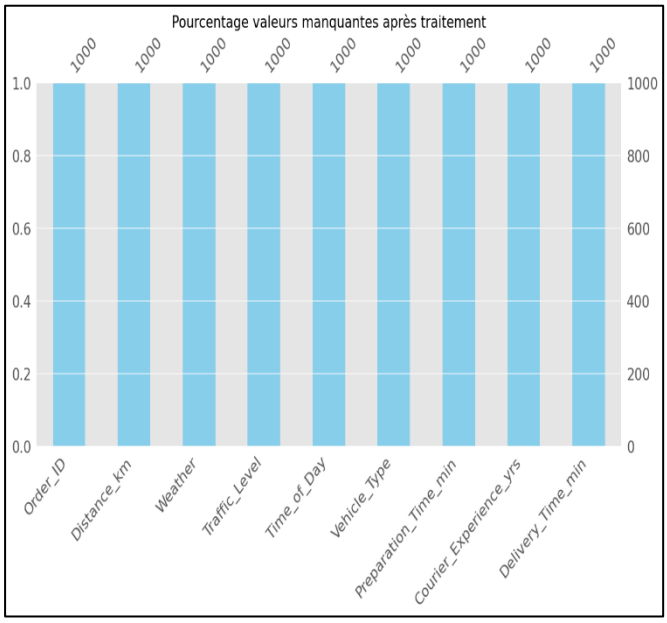


Figure 1 visualisations des valeurs manquantes

Les colonnes **Weather**, **Traffic_Level**, **Time_of_Day**, et **Courier_Experience_yrs** présentent chacune 30 valeurs manquantes. Ces variables sont essentielles pour l'analyse des performances de livraison, car elles peuvent affecter le temps de livraison. Il est donc crucial de traiter ces valeurs manquantes, par exemple par imputation, afin de garantir des analyses complètes et fiables.

b) Traitements des valeurs manquantes



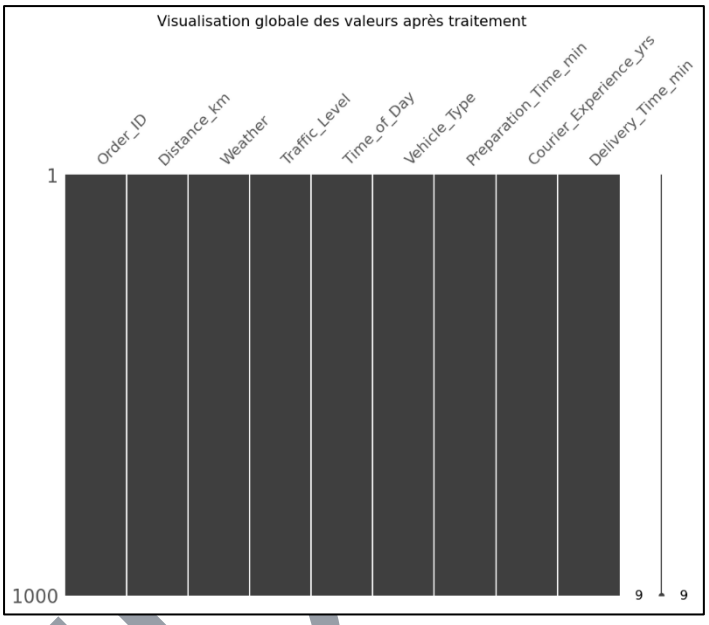


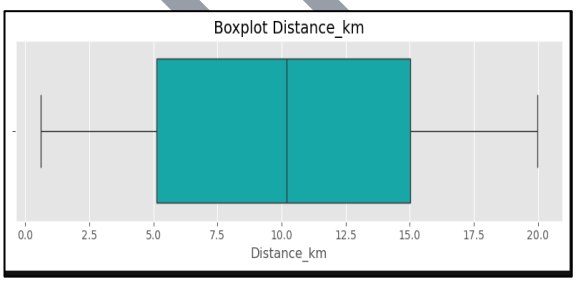
Figure 2 Traitements des valeurs manquantes (imputation)

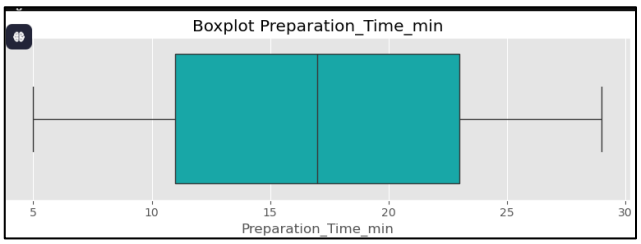
- **Observation** : Les barres sont toutes remplies, ce qui renforce l'idée que chaque variable maintenant ne contient aucune valeur manquante. Le traitement a donc permis d'assurer que toutes les informations requises sont présentes, favorisant une analyse ultime plus aisée.

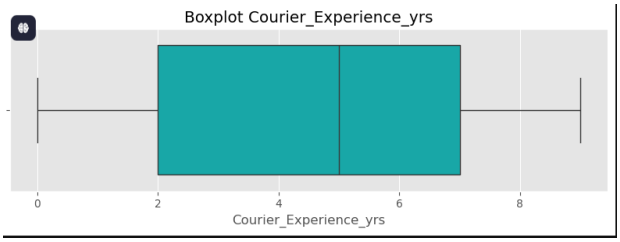
3.2) Visualisation et traitements des valeurs aberrantes / et ou Extrêmes

a) visualisation des valeurs manquantes

Dans ce cadre nous avons tenue compte des variables quantitative seulement







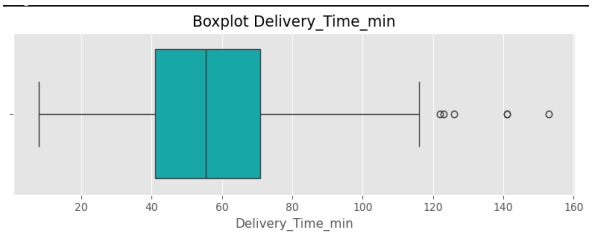


Figure 3 Visualisation des valeurs aberrantes

b) Traitement des valeurs aberrantes

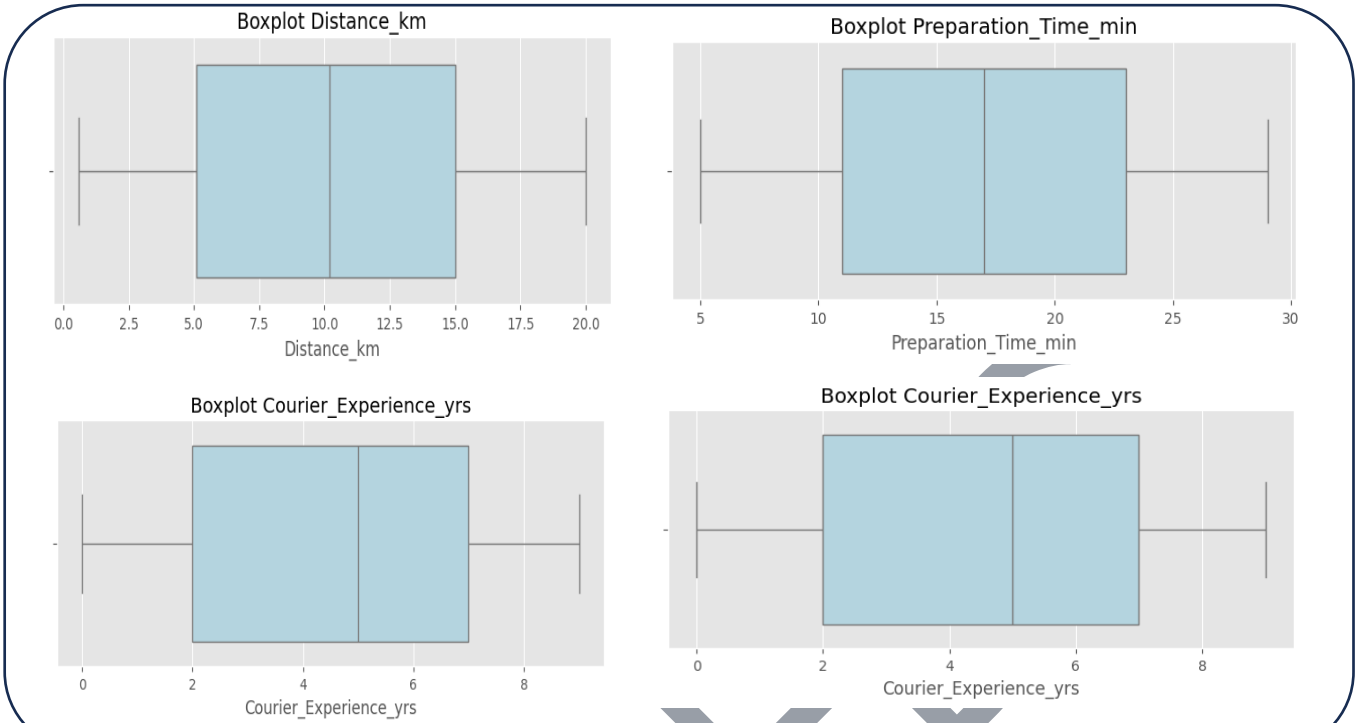


Figure 4 Traitement des valeurs manquantes

II) ANALYSE EXPLORATOIRE DES DONNEES (EDA)

L'analyse exploratoire des données (ou EDA, pour *Exploratory Data Analysis*) est une étape clé dans tout processus d'analyse de données. Son but principal est d'explorer et de mieux comprendre un jeu de données avant de passer à des modèles statistiques ou à des algorithmes complexes. Cette étape permet non seulement de formuler des hypothèses, mais aussi de repérer des tendances, des anomalies et des liens potentiels entre les variables. En plus de ça, elle est essentielle pour préparer les données en vue d'analyses plus poussées.

1) Analyses Univariées

1.1) Paramètres statistiques

Dans cette partie, nous allons aborder une analyse totalement univariée des variables. Nous allons cependant décrire quelques principales caractéristiques de nos variables quantitatives. Ce qui nous permettra d'avoir des détails plus instructifs avant d'entamer la modélisation statistique. Tableau des résumés numériques de nos variables. Ainsi que des graphiques.

Statistique	count	mean	std	min	25%	50% (median)	75%	max
Distance_km	1000	10.06	5.70	0.59	5.11	10.19	15.02	19.99
Preparation_Time_min	1000	16.98	7.20	5	11	17	23	29
Courier_Experience_yrs	1000	4.59	2.87	0	2	5	7	9
Delivery_Time_min	1000	56.62	21.71	8	41	55.5	71	116

Tableau 4 Paramètre statistiques

1.2) Interprétation des paramètres du tableau statistique

"Distance_km" et **"Courier_Experience_yrs"** :

Distance_km

- **Count** : 1000 observations. Cela reflète le nombre total de livraisons dans l'ensemble de données.
- **mean (moyenne)**: 10.06 km. En moyenne, les livraisons couvrent une distance de 10.06 km.
- **Std (écart-type)**: 5.70 km. L'écart-type relativement élevé indique que les distances sont assez dispersées autour de la moyenne, suggérant qu'il existe une grande variation dans les distances parcourues.
- **Min (minimum)**: 0.59 km. La distance la plus courte observée est de 0.59 km, ce qui pourrait suggérer des livraisons urbaines ou très locales.
- **25% (premier quartile)** : 5.11 km. 25 % des livraisons ont une distance inférieure ou égale à 5.11 km, ce qui indique une concentration notable de livraisons relativement courtes.
- **50% (médiane)** : 10.19 km. La médiane indique que la moitié des livraisons sont plus courtes que 10.19 km, et l'autre moitié est plus longue, ce qui est très proche de la moyenne.
- **75% (troisième quartile)** : 15.02 km. 75 % des livraisons ne dépassent pas 15.02 km, suggérant que des distances plus longues deviennent moins fréquentes.
- **Max (maximum)**: 19.99 km. La distance maximale observée dans les données est proche de 20 km, indiquant que les livraisons peuvent également couvrir de longues distances dans certains cas.

Interprétation : L'analyse des distances parcourues montre une grande variabilité. Bien que la plupart des livraisons se concentrent autour de la moyenne de 10 km, il existe des écarts notables, avec des livraisons aussi courtes que 0.59 km et aussi longues que 19.99 km. Cette variation pourrait refléter différentes zones géographiques desservies (urbaines vs. Rurales) ou des types de livraisons variés (petites vs. Grandes commandes).

Courier_Experience_yrs

- **Count** : 1000 individus. Cela indique que l'ensemble de données couvre 1000 coursiers.
- **mean (moyenne)**: 4.59 ans. L'expérience moyenne des coursiers est de 4.59 ans.
- **Std (écart-type)** : 2.87 ans. L'écart-type suggère une variation modérée dans l'expérience des coursiers.
- **Min (minimum)** : 0 ans. Certains coursiers n'ont aucune expérience préalable, ce qui peut poser un défi pour les formations.
- **25% (premier quartile)** : 0 ans. Un quart des coursiers sont débutants, sans aucune expérience, ce qui montre que de nombreux coursiers commencent leur carrière dans cette entreprise.
- **50% (médiane)** : 2 ans. La médiane est de 2 ans, ce qui signifie que la moitié des coursiers ont moins de 2 ans d'expérience. Cela suggère une main-d'œuvre relativement jeune en termes d'expérience.
- **75% (troisième quartile)**: 5 ans. 75 % des coursiers ont moins de 5 ans d'expérience. Cela indique une expérience principalement limitée à quelques années.
- **Max (maximum)** : 9 ans. Le coursier le plus expérimenté a 9 ans d'expérience.

Interprétation : L'analyse de l'expérience des coursiers révèle que la majorité d'entre eux ont relativement peu d'expérience, avec une médiane de seulement 2 ans. Un quart des coursiers sont totalement débutants, ce qui pourrait signifier qu'une formation continue ou un soutien supplémentaire serait nécessaire pour améliorer l'efficacité et la productivité de cette main-d'œuvre.

1.3) Graphiques des variables numériques

Nous visualiserons les graphiques sur lesquels les paramètres statistiques fussent effectuer.

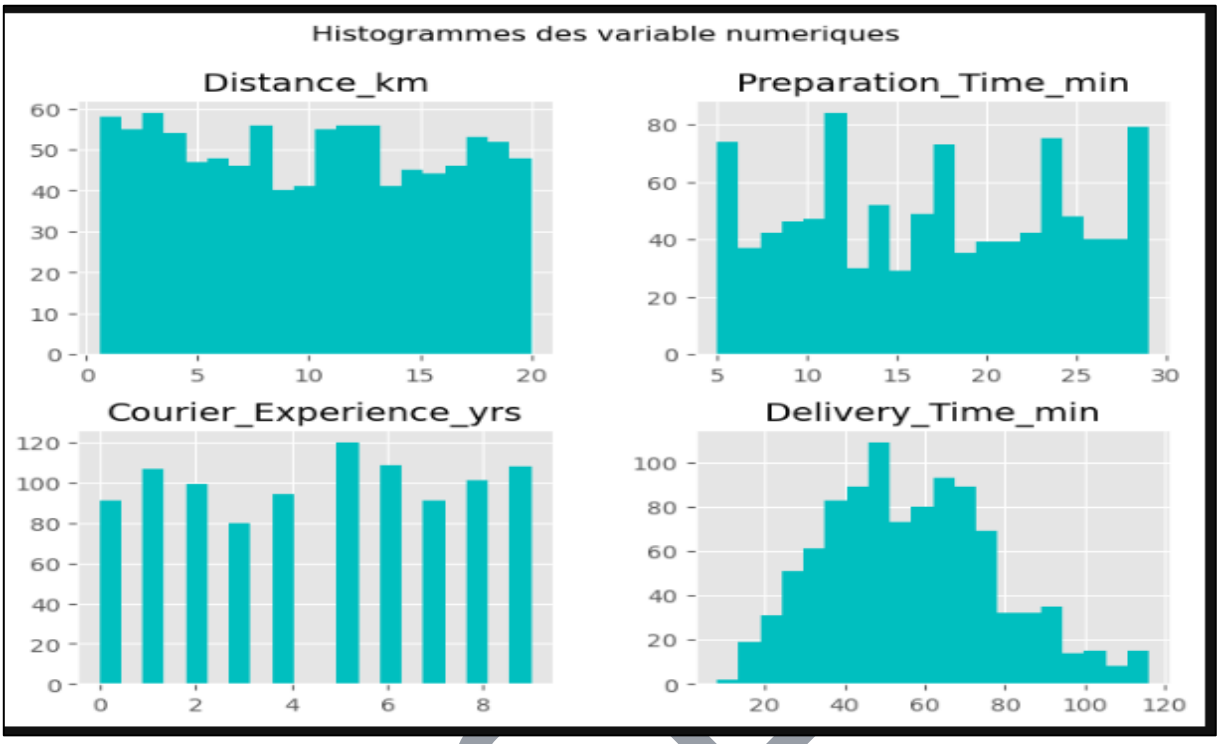


Figure 5 Variables numériques

1.4) Interprétation des graphiques des variables numeriques

1. Distance (km)

- **Distribution** : L'histogramme révèle une distribution assez uniforme avec quelques pics notables. La majorité des distances parcourues se situent entre 0 et 10 km, bien que certaines valeurs atteignent des distances plus longues, jusqu'à 50 km.
- **Interprétation** : Cette répartition suggère que la majorité des livraisons se déroulent sur de courtes distances, ce qui est typique pour les services de livraison urbains. Cela pourrait également indiquer une forte demande pour des livraisons locales ou intra-urbaines.

2. Temps de préparation (min)

- **Distribution** : L'histogramme montre une répartition assez variable, avec des valeurs allant de 0 à 30 minutes. Plusieurs pics sont observés, ce qui pourrait signifier que certains temps de préparation sont préférés ou standardisés.

- **Interprétation** : Cette variabilité suggère qu'il existe différents types de commandes ou de processus de préparation, certains nécessitant des délais plus longs que d'autres. Ces pics peuvent également indiquer des processus ou des pratiques d'efficacité spécifiques au sein de l'entreprise.

3. Expérience du coursier (ans)

- **Distribution** : L'histogramme montre une distribution assez équilibrée, avec une légère concentration des coursiers ayant entre 2 et 4 ans d'expérience.
- **Interprétation** : Cela indique une répartition relativement homogène des coursiers entre les nouveaux arrivants et ceux ayant quelques années d'expérience. Ce mélange peut influencer l'efficacité générale des livraisons, en apportant à la fois des perspectives nouvelles et des connaissances pratiques sur le terrain.

4. Temps de livraison (min)

- **Distribution** : L'histogramme présente une distribution asymétrique, avec une concentration notable autour des 20 à 30 minutes, mais aussi des cas extrêmes allant jusqu'à 120 minutes.
- **Interprétation** : Cela montre que la plupart des livraisons sont effectuées rapidement, mais qu'il existe aussi des cas où des retards importants se produisent. Ces anomalies pourraient être attribuées à des facteurs externes (météo, circulation) ou à des problèmes logistiques internes (préparation, gestion des ressources).

Conclusion

Les graphiques offrent une vue d'ensemble précieuse des opérations de livraison, mettant en évidence des tendances importantes comme la concentration des distances et des temps de livraison. Ces observations peuvent être utilisées pour affiner les processus opérationnels, ajuster la répartition des ressources (par exemple, en fonction des distances et des temps de préparation), et identifier les facteurs contribuant aux retards. Une analyse plus approfondie des corrélations entre ces variables pourrait fournir des informations supplémentaires sur les performances opérationnelles et aider à optimiser davantage les processus de livraison.

2) Analyse Bivariees

2.1) Matrice de corrélation

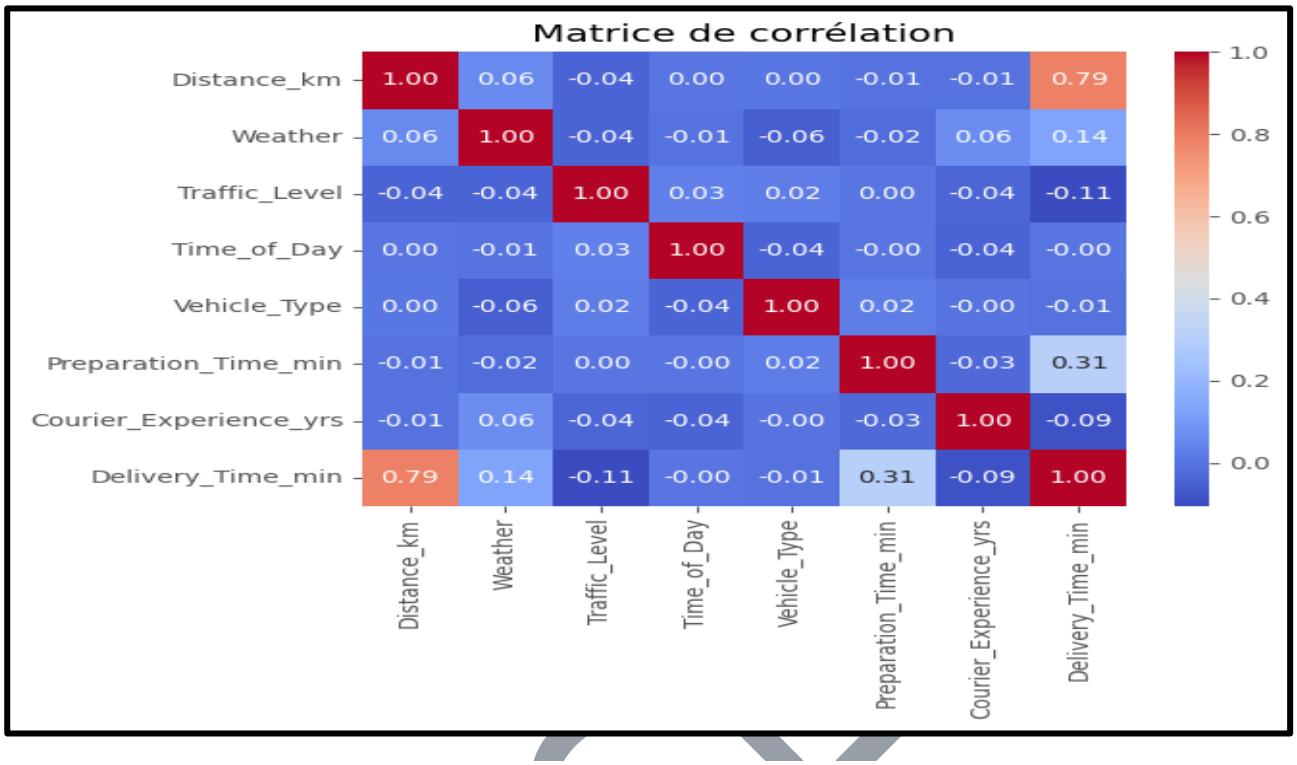


Figure 6 Matrice de corrélation

Distance (km)

- **Corrélations positives** :
  - **Delivery Time (min)** : Le coefficient de 0.79 montre une forte corrélation positive, indiquant que plus la distance est longue, plus le temps de livraison tend à être élevé. Cela est attendu, car les livraisons sur de longues distances prennent généralement plus de temps en raison de la distance à parcourir.
- **Corrélations faibles** :
  - **Preparation Time (min)** : Le coefficient de 0.31 suggère qu'il existe une légère tendance à ce que des distances plus longues soient associées à des temps de préparation plus longs. Cependant, cette relation est faible, ce qui signifie qu'elle n'a pas une influence marquée sur le temps de préparation.

Preparation Time (min)

- **Corrélations positives** :
  - **Delivery Time (min)** : Le coefficient de 0.31 montre une faible corrélation positive, indiquant qu'un temps de préparation plus long est légèrement associé à un temps de livraison plus long. Toutefois, cette relation est assez faible et nécessite une exploration plus approfondie pour confirmer l'impact.

- **Corrélations négatives** :
    - **Weather** et **Traffic Level** : Ces variables montrent des corrélations faibles, ce qui suggère que les conditions météorologiques et le niveau de trafic n'ont pas un impact significatif sur le temps de préparation des livraisons.

Courier Experience (yrs)

- **Corrélations négatives** :
    - **Delivery Time (min)** : Le coefficient de -0.09 indique qu'il n'y a pratiquement aucune corrélation entre l'expérience du coursier et le temps de livraison. Cela suggère que l'expérience des coursiers n'affecte pas de manière notable la durée des livraisons.
- **Corrélations faibles avec d'autres variables** : Les faibles corrélations avec d'autres variables laissent entendre que l'expérience des coursiers n'a pas d'impact majeur sur les performances de livraison, ni sur les autres aspects comme le temps de préparation.

Weather, Traffic Level, et Time of Day

Ces trois variables présentent des corrélations très faibles avec les autres, ce qui signifie qu'elles n'ont pas d'influence significative sur les temps de préparation ni sur les temps de livraison. Cela suggère que, dans ce cas précis, ces facteurs externes n'affectent pas de manière notable les performances des livraisons.

3. Conclusions et recommandations

- **Optimisation des distances** : La forte corrélation entre la distance et le temps de livraison (0.79) indique qu'il serait pertinent d'explorer des solutions pour optimiser les itinéraires, comme la planification de trajets plus courts ou l'amélioration des processus de logistique, afin de réduire les délais de livraison.
- **Évaluation des temps de préparation** : Bien qu'il existe une corrélation faible entre le temps de préparation et d'autres variables, il pourrait être bénéfique d'examiner plus en détail les processus de préparation. En identifiant les facteurs influençant les pics ou les variations du temps de préparation, l'entreprise pourrait trouver des leviers d'amélioration pour rendre le processus plus efficace.
- **Formation des courriers** : Avec une faible corrélation entre l'expérience des coursiers et le temps de livraison, il pourrait être judicieux d'investir davantage dans la formation des courriers. D'autres facteurs, tels que la gestion du temps ou l'efficacité dans la navigation, pourraient également être plus déterminants dans les performances de livraison que l'expérience en elle-même.

PARTIE II : REGRESSION LINEAIRE MULTIPLES

La régression linéaire multiple est une technique statistique qui permet de modéliser la relation entre une variable dépendante (ou à expliquer) et plusieurs variables indépendantes (ou explicatives). Contrairement à la régression linéaire simple, qui n'implique qu'une seule variable explicative, la régression linéaire multiple prend en compte plusieurs facteurs pouvant influencer la variable à expliquer.

Formule générale de la régression linéaire multiple :

La forme générale d'un modèle de régression linéaire multiple est la suivante :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

- Y : Variable dépendante (à expliquer)
- β_0 : Ordonnée à l'origine (intercept)
- $\beta_1, \beta_2, \ldots, \beta_p$: Coefficients des variables indépendantes (pentes)
- $X_1, X_2, \ldots, X_p$: Variables indépendantes (explicatives)
- ε : Terme d'erreur (erreur aléatoire)

Équation 1 Formule de régressions linéaires multiples

Objectif :

Le but de la régression linéaire multiple est d'estimer les coefficients β1, β2, βp\beta_1, \beta_2, \dots, \beta_pβ1, β2,…,βp qui minimisent l'écart entre les valeurs prédites par le modèle et les valeurs réelles de la variable dépendante YYY. Cela se fait généralement par la méthode des moindres carrés, qui minimise la somme des carrés des résidus.

1) Construction du modèle de régression linéaire multiple

La **construction d'un modèle de régression** est un processus d'analyse statistique visant à comprendre la relation entre une variable dépendante (ou variable cible) et une ou plusieurs variables indépendantes (ou explicatives). L'objectif est de créer un modèle qui peut prédire ou expliquer les valeurs de la variable dépendante en fonction des valeurs des variables indépendantes.

1.1) Partitionnement des données en ensemble d'entraînement et de test

Le cas de partitionnement du model se fait plus dans les cas de machine Learning mais nous pour notre modelé nous utiliseront une alternative externe et efficace.

Ensemble	Nombre d'exemples	Nombre de caractéristiques
Ensemble d'entraînement	800	2
Ensemble de test	200	2

L'ensemble d'entraînement contient 800 exemples, chacun avec 2 caractéristiques.

L'ensemble de test contient 200 exemples, également avec 2 caractéristiques.

Tableau 5 Partitionnement

1.2) Représentation des variables

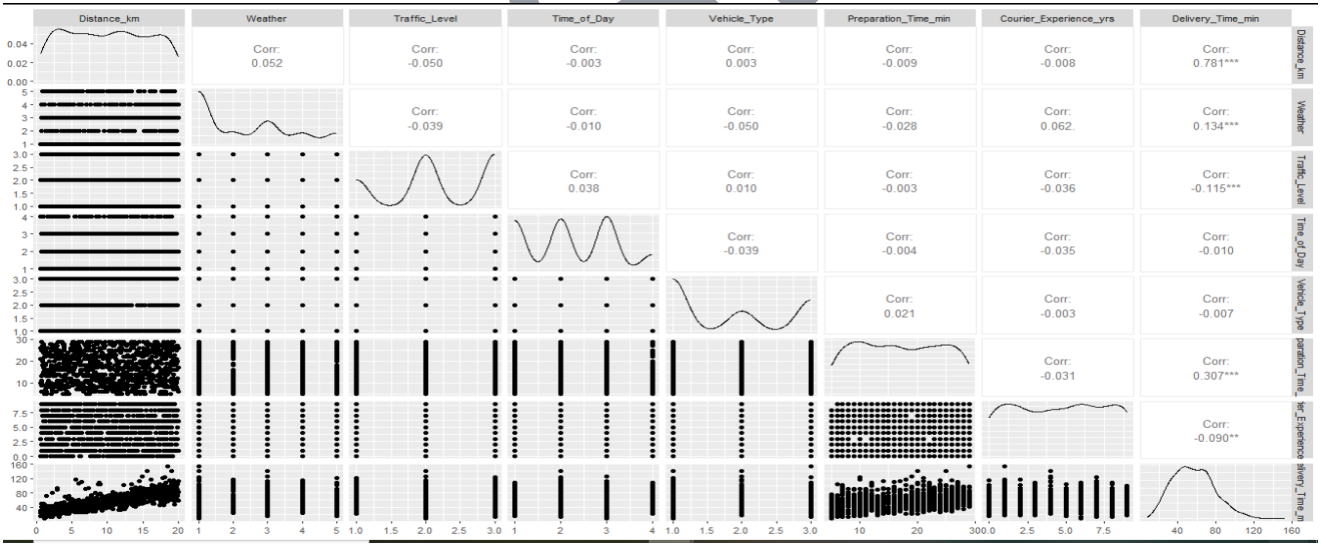


Figure 7 Représentation des variables

Distance_km : 0.787378

- Ce coefficient positif élevé indique une forte corrélation positive entre la distance parcourue et le temps de livraison. Autrement dit, plus la distance est grande, plus le temps de livraison a tendance à être long.

Traffic_Level : -0.115355

- Ce coefficient négatif indique une faible corrélation négative entre le niveau de trafic et le temps de livraison. Cela signifie que lorsque le trafic est plus élevé, le temps de livraison a tendance à être légèrement plus court.

Vehicle_Type : -0.008590

- Ce coefficient négatif très faible suggère qu'il n'y a pratiquement pas de corrélation entre le type de véhicule utilisé et le temps de livraison.

Courier_Experience_yrs : -0.088108

- Ce coefficient négatif indique une faible corrélation négative entre l'expérience du livreur et le temps de livraison. Autrement dit, plus l'expérience du livreur est grande, plus le temps de livraison a tendance à être court.

1.3) Modèle de régressions linéaires

Variable	Coefficient
Distance_km	17.209173
Preparation_Time_min	7.001535

Figure 8 Modèle de régression

- **Distance_km,(17.209173)** :
  Cela signifie que pour chaque augmentation d'1 kilomètre dans la distance (en supposant que c'est une relation linéaire), la variable dépendante (que ce soit un coût, un temps, une estimation, etc.) augmente de 17.21 unités. Par exemple, si la variable dépendante est un coût, cela pourrait signifier qu'augmenter la distance de 1 km augmenterait le coût de 17.21 unités monétaires.
- **Preparation_Time_min,(7.001535)** :
  Cela indique que pour chaque minute supplémentaire dans le temps de préparation, la variable dépendante augmente de 7.00 unités. Si la variable dépendante est, par exemple, un prix, chaque minute supplémentaire de préparation entraînera une augmentation de 7 unités.

1.4) Estimation des paramètres du modèle

L'**estimation des paramètres du modèle** fait référence au processus qui permet de déterminer les valeurs des paramètres (ou coefficients) d'un modèle statistique, comme le modèle de régression linéaire. Dans ce contexte, l'objectif est de trouver les valeurs des paramètres qui permettent au modèle de mieux s'ajuster aux données observées.

1.4.1) estimation des coefficients

Variable	Coefficient
(Intercept)	15.2112556
Distance_km	2.9662356
Preparation_Time_min	0.9445546
Courier_Experience_yrs	-0.5537925
Weather	1.6104985
Traffic_Level	-1.9803384
Time_of_Day	-0.2065987
Vehicle_Type	-0.4903394

Figure 9 Estimation des coefficients

Explication des coefficients :

(Intercept) : La valeur de la variable dépendante lorsque toutes les variables indépendantes sont égales à zéro (dans ce cas, 15.21).

Distance_km : Chaque kilomètre supplémentaire augmente la variable dépendante de 2.97 unités.

Preparation_Time_min : Chaque minute supplémentaire de préparation augmente la variable dépendante de 0.94 unités.

Courier_Experience_yrs : Chaque année supplémentaire d'expérience de coursier diminue la variable dépendante de 0.55 unités.

Weather : Une augmentation de l'index météo (qui peut être une mesure comme l'indice de chaleur ou l'humidité) augmente la variable dépendante de 1.61 unités.

Traffic_Level : Un niveau de trafic plus élevé diminue la variable dépendante de 1.98 unités.

Time_of_Day : Le moment de la journée affecte négativement la variable dépendante de 0.21 unités.

Vehicle_Type : Le type de véhicule affecte négativement la variable dépendante de 0.49 unités.

1.4.2) Visualisation de l'intervalle de confiance des coefficients estimés

Variable	2.5%	97.5%
(Intercept)	10.9077761	19.5147351
Distance_km	2.8362598	3.0962114
Preparation_Time_min	0.8428501	1.0462591
Courier_Experience_yrs	-0.8071732	-0.3004119
Weather	1.0740796	2.1469173
Traffic_Level	-2.9613754	-0.9993013
Time_of_Day	-0.9789577	0.5657602
Vehicle_Type	-1.3382613	0.3575825

Figure 10intervalle de confiance

(Intercept) : La valeur de la variable cible varie entre **10.91** et **19.51** lorsque toutes les autres variables sont nulles.

Distance_km : Chaque kilomètre supplémentaire augmente la variable cible de **2.84 à 3.10** unités.

Preparation_Time_min : Chaque minute de préparation augmente la variable cible de **0.84 à 1.05** unités.

Courier_Experience_yrs : Plus d'expérience diminue la variable cible de **0.81 à 0.30** unités.

Weather : L'impact de la météo varie entre **1.07 et 2.15** unités, suggérant un effet positif.

Traffic_Level : Plus de trafic diminue la variable cible de **2.96 à 1.00** unités.

Time_of_Day : L'heure de la journée a un effet modéré, variant de **-0.98 à 0.57**.

Vehicle_Type : Le type de véhicule influence la variable cible de **-1.34 à 0.36** unités, suggérant un impact négatif variable.

1.4.3) résultat final de l'estimation

```
Residuals:
    Min      1Q  Median      3Q     Max
-25.786  -6.292  -0.803   4.483  67.541

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              15.21126    2.19266   6.937 7.77e-12 ***
Distance_km               2.96624    0.06622  44.791  < 2e-16 ***
Preparation_Time_min      0.94455    0.05182  18.228  < 2e-16 ***
Courier_Experience_yrs   -0.55379    0.12910  -4.290 1.99e-05 ***
Weather                   1.61050    0.27331   5.893 5.42e-09 ***
Traffic_Level            -1.98034    0.49985  -3.962 8.04e-05 ***
Time_of_Day              -0.20660    0.39352  -0.525    0.600
Vehicle_Type             -0.49034    0.43202  -1.135    0.257
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.16 on 875 degrees of freedom
  (117 observations effacées parce que manquantes)
Multiple R-squared:  0.7345,    Adjusted R-squared:  0.7324
F-statistic: 345.8 on 7 and 875 DF,  p-value: < 2.2e-16
```

Figure 11 Results_estimation

Delivery_time = 15.21126+2.96624 **Dkm**+0.94455 **Prep_Time_min**−0.55379 **C_E_y**+1.61050Weather−1.98034 **Traffic**

Ce modèle de régression multiple présente un bon ajustement des données. Les coefficients sont statistiquement significatifs et le modèle explique une proportion importante de la variabilité des données, comme l'indiquent les valeurs de R-carré élevées. Les statistiques des résidus sont également satisfaisantes.

☐ Temps **de livraison** de base (sans aucune variable) est de 15.21126 unités.

☐ Plus **la distance augmente**, plus le temps de livraison augmente (coefficient positif pour Dkm).

☐ Plus **le temps de préparation est long**, plus le temps de livraison augmente également.

☐ Les **années qui passent semblent réduire le temps de livraison**, peut-être grâce à l'optimisation des processus.

☐ Les **conditions météorologiques défavorables** (si elles sont mesurées comme une variable continue) augmentent le temps de livraison.

☐ Le **trafic a un impact surprenant** : une augmentation du trafic semble réduire le temps de livraison selon cette équation.

1.4.5) choix des variables

Nb : Apres détection des variables une, une nouvelle régression a été effectué sur les variables

```
Residuals:
    Min      1Q  Median      3Q     Max
-25.300  -6.417  -0.661   4.454  69.831

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               10.89487    1.73590   6.276 5.30e-10 ***
henoc$Distance_km          2.99417    0.06564  45.617  < 2e-16 ***
henoc$Preparation_Time_min 0.97260    0.05147  18.896  < 2e-16 ***
henoc$Traffic_Level       -2.06334    0.49682  -4.153 3.58e-05 ***
henoc$Weather              1.60640    0.26869   5.979 3.20e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.42 on 935 degrees of freedom
  (60 observations effacées parce que manquantes)
Multiple R-squared:  0.7294,     Adjusted R-squared:  0.7283
F-statistic: 630.2 on 4 and 935 DF,  p-value: < 2.2e-16
```

Figure 12choix des variables

Interprétation du choix des variables dans le modèle de régression :

1. **Variables explicatives retenues :**

henocsDistance_km : La distance parcourue est une variable clé pour expliquer la variabilité des résidus.

henoc$Preparation_Time_min : Le temps de préparation a également été intégré au modèle.

henoc$Traffic_Level : Le niveau de trafic a été pris en compte.

henocsweather : Les conditions météorologiques influencent également les résidus.

2. **Justification du choix des variables :**

Ces variables représentent des facteurs déterminants qui peuvent affecter les performances du système étudié, tels que la distance, le temps de préparation, le trafic et les conditions météorologiques.

Leur inclusion dans le modèle permet de mieux expliquer la variabilité des résidus, ce qui se reflète dans un **R-carré élevé** (0.7294), indiquant une bonne capacité du modèle à expliquer la variation des données.

Toutes les variables ont des coefficients statistiquement significatifs, ce qui montre qu'elles contribuent de manière importante à la précision du modèle.

1.5. Calcul des métriques de performance du modèle

Une **métrique de performance** est une mesure quantifiable utilisée pour évaluer l'efficacité ou la réussite d'un modèle, d'un processus ou d'un système. Dans notre contexte des modèles de régression, les métriques de performance permettent de juger de la qualité du modèle par rapport à sa capacité à prédire ou à expliquer les données.

Métriques	Train	Test
R^2	0.715340	0.734336
MSE	135.760156	118.064637
RMSE	11.651616	10.865755

Métriques	Valeurs
Durbin-Watson	2.007057
Jarque-Bera Statistic	2140.803291
Jarque-Bera p-value	0.000000
Skewness	1.864054
Kurtosis	7.094033
Condition Number	1.016810

Métriques	Valeurs
F-statistic	1001.415844
Prob (F-statistic)	0.000000
Log-Likelihood	-3099.506735
AIC	3934.711817
BIC	3948.765653
n	800.000000
Df Residuals	797.000000
Df_model	2.000000

Réel	Prédiction
521	41.436166
737	65.678281
740	35.708194
660	43.266346
411	84.731093
...	...
408	79.257667
332	28.971670
208	56.445634
613	44.975235
78	42.687966

Figure 13 Métriques de performances

❖ ***prédiction du temps de livraison***

Temps de livraison prédit
54.60149830343386 minutes

Le **temps de livraison prédit** est de **54.60 minutes**. Cela signifie que le modèle estime qu'il faudra environ **54 minutes et 36 secondes** pour effectuer la livraison.

2) Analyse Résiduels et Validation du Modèle de Régression

2.1) Analyse des résidus

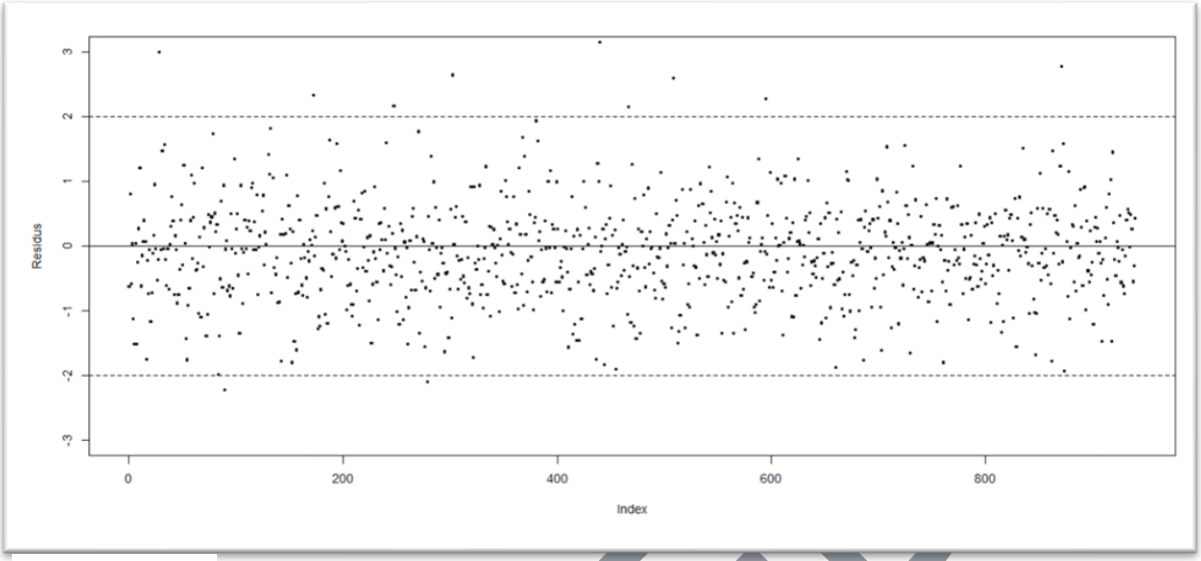


Figure 14 Résidus

Observations Clés :

1. **Distribution des Résidus** : Les résidus sont répartis de manière relativement homogène sur l'ensemble de l'index, sans forme évidente de tendance systématique. Cette absence de structure ou de pattern dans les résidus suggère que le modèle utilisé pour les prédictions ne souffre pas d'un biais systématique majeur.
2. **Amplitude des Résidus** : Les résidus varient entre environ -3 et 1, ce qui indique que le modèle fait globalement des prédictions précises. Cependant, quelques résidus atteignent des valeurs relativement élevées (en valeur absolue), ce qui pourrait signaler que certaines observations posent plus de difficultés au modèle, et donc méritent une attention particulière.
3. **Homogénéité de la Variance (Homoscedasticité)** : La dispersion des résidus est homogène tout au long de l'index, sans signes évidents d'hétéroscédasticité (variance non constante). Cela suggère que les hypothèses du modèle, en particulier l'hypothèse d'homogénéité de la variance, sont globalement respectées, ce qui est un indicateur positif pour la validité du modèle.

2.2) Validation du Modèle de Régression

Dans cette partie, nous effectuerons certains tests pour vérifier l'efficacité du modèle

a) Test de linéarité

Hypothèse **:**

- - H0 : Le modèle est linéaire
- - H1 : Le modèle est non linéaire

```
        Rainbow test

data:  reg.fin
Rain = 1.3773, df1 = 470, df2 = 465, p-value = 0.0002799
```

Figure 15 Linéarité

La p-value est inférieure à 0.05 on rejette H0. Autrement dire, le modèle obtenu n'est pas un modèle linéaire

b) Test de multi colinéarité

Le test de multi colinéarité se fait avec le calcul des VIF (variance inflation factor). Les VIF estiment de combien la variance d'un coefficient est « augmentée » en raison d'une relation linéaire avec d'autres prédicteurs.

Hypothèse :

- Si tous VIF sont supérieurs à 1 alors les prédicteurs sont colorés
- Si tous VIF sont égaux à 1 alors il existe une multi colinéarité.

Variables	Tolerance	VIF
henoc$Distance_km	0.9940320	1.006004
henoc$Preparation_Time_min	0.9990234	1.000978
henoc$Traffic_Level	0.9965454	1.003467
henoc$Weather	0.9942560	1.005777

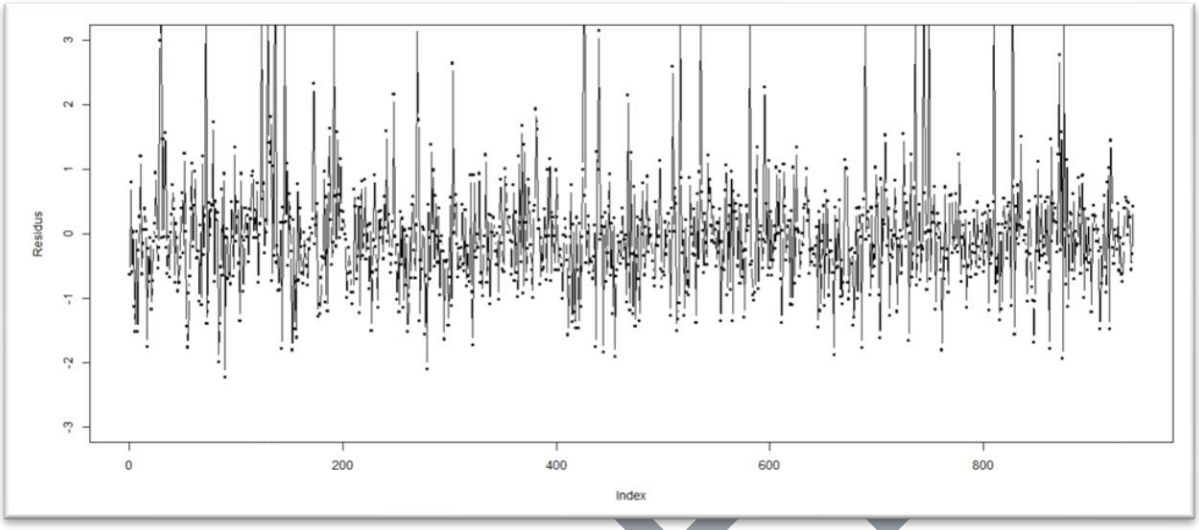
Figure 16 Multi colinéarité

Conclusion D'après le résultat, tous les VIF sont proches de 1 il n'y a donc pas de problème potentiel de colinéarité à explorer dans notre modèle.

c) Test de l'autocorrélation des erreurs

Hypothèse :

- H0 : Il y a corrélation des erreurs
- H1 : I n'y a aucune corrélation entre les erreurs




```
> acf(residus,plot = FALSE)

Autocorrelations of series 'residus', by lag

    0     1     2     3     4     5     6     7     8     9    10    11    12    13
1.000 0.020 0.027 -0.026 0.023 0.000 -0.040 0.039 0.019 0.038 -0.001 -0.019 -0.024 0.017
   14    15    16    17    18    19    20    21    22    23    24    25    26    27
0.019 -0.037 -0.033 0.037 0.014 0.046 0.014 -0.043 0.011 -0.030 -0.002 -0.072 -0.018 0.033
   28    29
0.029 -0.041
```

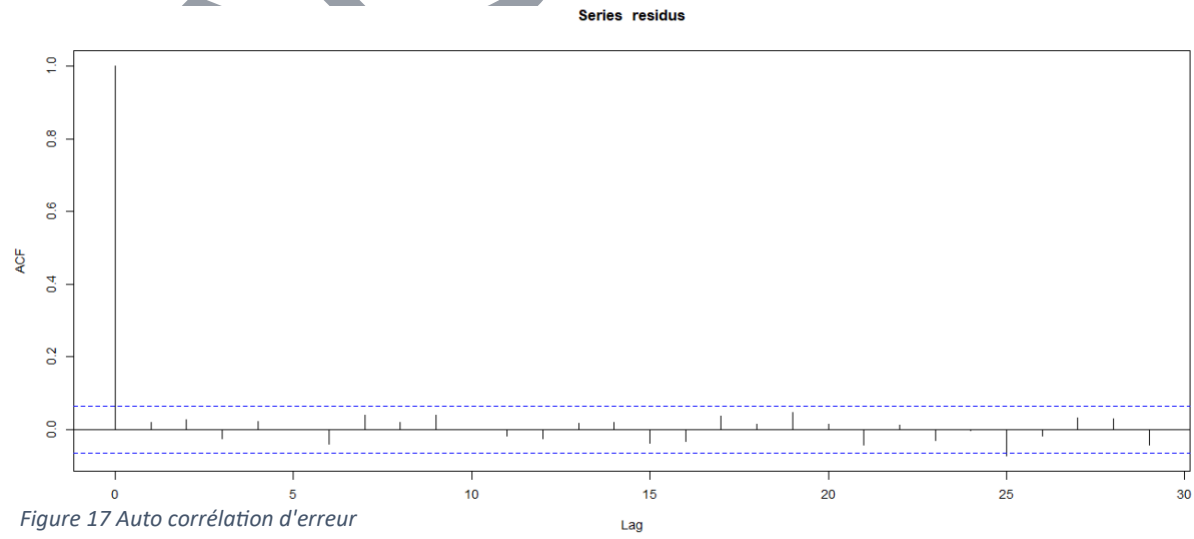


Figure 17 Auto corrélation d'erreur

d) Test d'homoscédasticité des erreurs

Hypothèse :

- H0 : Il y a homoscédasticité
- H1 : Il y a hétéroscédasticité

Test	BP	df	p-value
Breusch-Pagan Test (studentized)	9.7277	4	0.04527

La p-value est inférieure à 0,05. On ne peut admettre H0 autrement dire, il y a hétéroscédasticité des erreurs. Cela se visualise sur les graphes ci-dessous

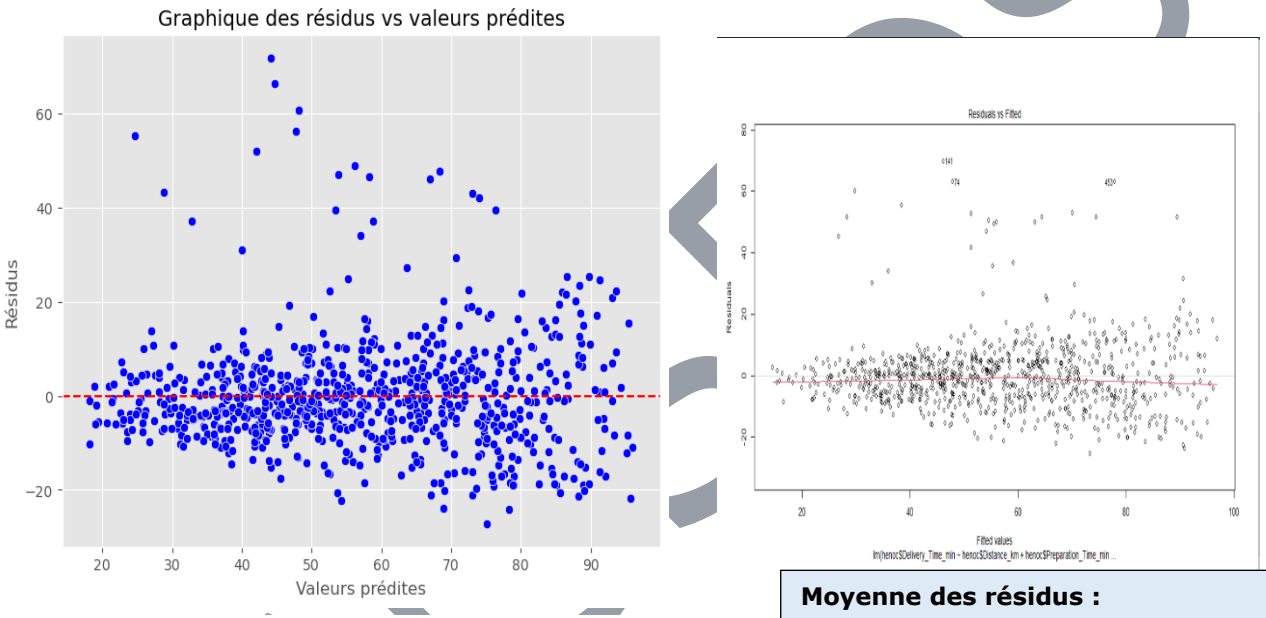


Figure 18 RESIDUS VS VALEURS PREDITES

e) Moyenne des termes d'erreur

```
> mean(residus)
[1] -3.082221e-16
```

Moyenne des résidus :

La moyenne des résidus est de -3.082221e-16, soit très proche de 0.

Cette valeur proche de zéro indique que le modèle de régression semble bien centrer les résidus autour de la moyenne, ce qui est un bon signe.

Test de normalité de Shapiro-Wilk :

Le test de Shapiro-Wilk est utilisé pour vérifier si les résidus suivent une distribution normale.

La statistique W du test est de 0.84755, avec une p-value inférieure à 2.2e-16.

Comme la p-value est très faible (< 0.05), on peut rejeter l'hypothèse de normalité des résidus.

Cela signifie que les résidus ne suivent pas une distribution normale, ce qui peut indiquer des problèmes dans les hypothèses du modèle de régression.

f) Test normalité des erreurs

```
        Shapiro-Wilk normality test

data:  residus
W = 0.84755, p-value < 2.2e-16
```

3) Prédiction d'une nouvelle valeur

```
Call:
lm(formula = henoc.Delivery_Time_min ~ henoc.Distance_km + henoc.Preparation_Time_min +
    henoc.Traffic_Level + henoc.Weather, data = henoc1)

Residuals:
    Min      1Q  Median      3Q     Max
-20.478  -5.107  -1.219   3.327  64.365

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                14.33591    1.32527  10.817  < 2e-16 ***
henoc.Distance_km           2.96141    0.06028  49.126  < 2e-16 ***
henoc.Preparation_Time_min  0.95955    0.04718  20.336  < 2e-16 ***
henoc.Traffic_LevelLow    -11.87703    0.93235 -12.739  < 2e-16 ***
henoc.Traffic_LevelMedium  -5.92417    0.92826  -6.382 2.83e-10 ***
henoc.WeatherFoggy          7.97890    1.13871   7.007 4.87e-12 ***
henoc.WeatherRainy          4.72432    0.89019   5.307 1.41e-07 ***
henoc.WeatherSnowy          9.92914    1.20453   8.243 6.11e-16 ***
henoc.WeatherWindy          1.87860    1.20373   1.561    0.119
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.14 on 874 degrees of freedom
Multiple R-squared:  0.7809,    Adjusted R-squared:  0.7789
F-statistic: 389.4 on 8 and 874 DF,  p-value: < 2.2e-16
```

```
> Xnew
  henoc.Distance_km henoc.Preparation_Time_min henoc.Traffic_Level2
1                45                          1                    0
> |
```

Figure 19 Analyse prévisionnelle

```
> print(predictions)
       fit      lwr      upr
1 142.6349 122.2074 163.0624
```

Après avoir exploré la régression linéaire et ses capacités à modéliser la relation entre une variable dépendante continue et des variables indépendantes, il est pertinent de se tourner vers des méthodes statistiques qui permettent d'analyser des situations où la variable dépendante est influencée par des facteurs catégoriels ou qualitatifs. C'est dans ce contexte que l'ANOVA (analyse de la variance) prend tout son sens. Contrairement à la régression linéaire, qui est principalement utilisée pour examiner l'effet de variables continues, l'ANOVA permet d'étudier si les moyennes de plusieurs groupes, définis par des facteurs catégoriels, diffèrent de manière significative. Cette approche nous permet ainsi de mieux comprendre les effets des différentes catégories d'une variable indépendante sur la variable dépendante.

Ainsi, tout en maintenant une approche axée sur l'analyse de l'impact des variables explicatives sur notre variable cible, nous allons maintenant explorer l'ANOVA pour tester des hypothèses liées aux différences de moyennes entre les groupes, et en particulier la manière dont ces différences peuvent expliquer la variation des résultats observés.

PARTIE III : ANOVA

L'**ANOVA** (Analyse de la Variance) est une méthode statistique utilisée pour tester si les moyennes de plusieurs groupes ou échantillons sont significativement différentes les unes des autres. En d'autres termes, elle permet de déterminer si les différences observées entre les groupes sont dues à des variations réelles ou simplement à des fluctuations aléatoires.

Principe de l'ANOVA :

L'ANOVA compare la **variance** (les écarts par rapport à la moyenne) à l'intérieur de chaque groupe et la **variance** entre les groupes. Si la variance entre les groupes est significativement plus grande que la variance au sein des groupes, cela suggère qu'il existe des différences significatives entre les moyennes des groupes.

1) Graphe des délais de livraison selon la météo

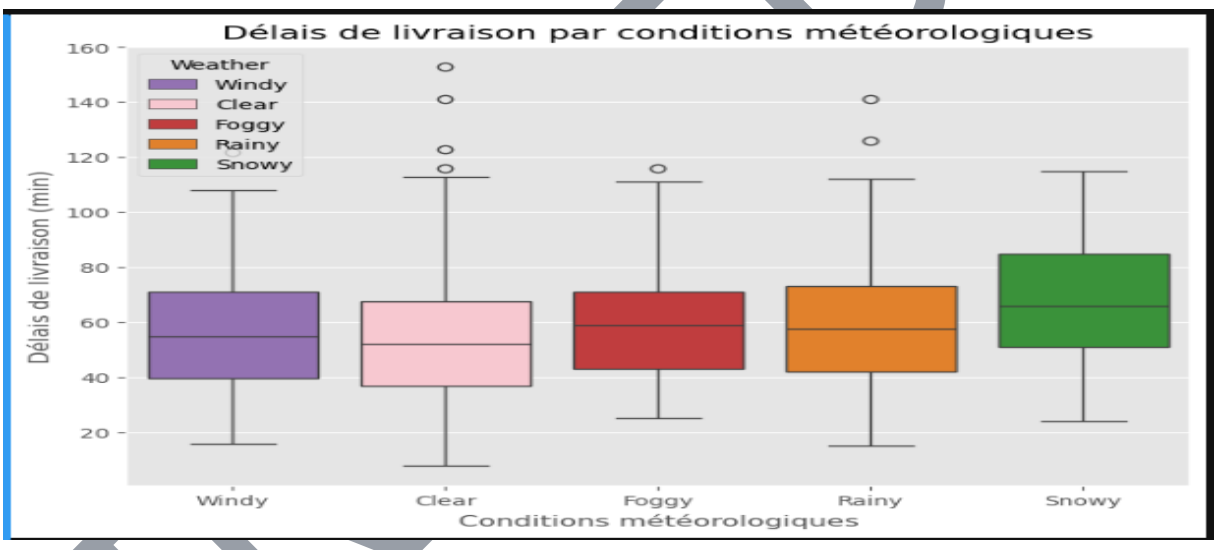


Figure 20 Délais de livraison en fonction de la météo

2) Estimation des statistiques de base (mean, quantile, sd)

```
> tapply(henoc$Delivery_Time_min,henoc$Weather,mean,na.rm=TRUE)
   Clear    Foggy    Rainy    Snowy    Windy
53.08298 59.46602 59.79412 67.11340 55.45833
> tapply(henoc$Delivery_Time_min,henoc$Weather,sd,na.rm=TRUE)
   Clear    Foggy    Rainy    Snowy    Windy
21.27221 20.86221 22.82244 21.29157 21.77779
> tapply(henoc$Delivery_Time_min,henoc$Weather,quantile,na.rm=TRUE)
$Clear
    0%    25%    50%    75%   100%
  8.00  37.00  52.00  67.75 153.00
```

Tableau 6 Estimation statistiques

3) Teste de la normalité des données dans chaque temps météorologique.

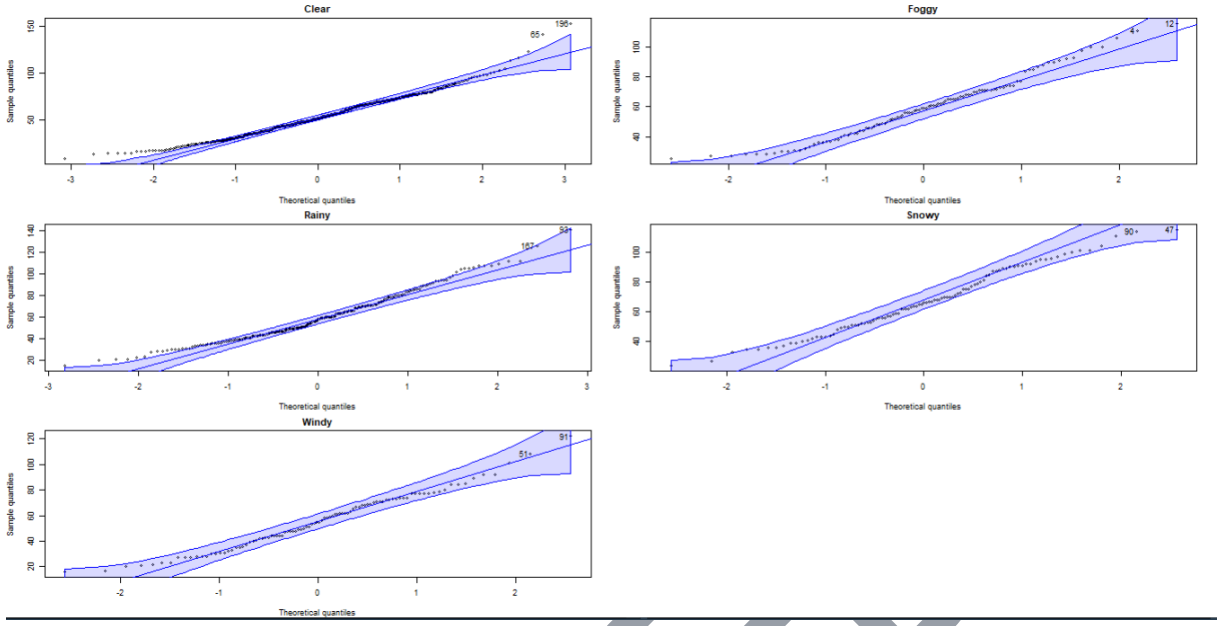


Figure 21 représentation des temps

3.1) Normalité

```
        Shapiro-Wilk normality tests

data:  henoc$Delivery_Time_min by henoc$Weather

           W   p-value
Clear 0.9761 5.916e-07 ***
Foggy 0.9717 0.0261689 *
Rainy 0.9672 0.0001105 ***
Snowy 0.9800 0.1467843
Windy 0.9786 0.1190879
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tableau 7 Shapiro test

1. Interprétation

- Pour la plupart des conditions météorologiques, la p-value est inférieure à 0,05, indiquant que l'hypothèse de normalité des données peut être rejetée.
- Seule la condition "Snowy" a une p-value supérieure à 0,05, suggérant que les données pour cette condition pourraient suivre une distribution normale.
- Les codes de signification indiquent que les p-values inférieures à 0,001 sont considérées comme "très significatives" (***), celles inférieures à 0,01 sont "très significatives" ()***, et celles inférieures à 0,05 sont "significatives" ().

4) Teste de l'égalité des variances

```
> bartlett.test(henoc$Delivery_Time_min~henoc$Weather)

        Bartlett test of homogeneity of variances

data:  henoc$Delivery_Time_min by henoc$Weather
Bartlett's K-squared = 1.7803, df = 4, p-value = 0.7761

> kruskal.test(henoc$Delivery_Time_min~henoc$Weather)

        Kruskal-Wallis rank sum test

data:  henoc$Delivery_Time_min by henoc$Weather
Kruskal-Wallis chi-squared = 36.768, df = 4, p-value = 2.011e-07
```

1. Test de Bartlett : Test d'homogénéité des variances

Le test de Bartlett est utilisé pour tester si les variances sont égales entre plusieurs groupes. C'est un test paramétrique qui suppose que les données suivent une distribution normale dans chaque groupe.

- **Résultat du test de Bartlett** :
  - **K-squared = 1.7803**
  - **df = 4**
  - **p-value = 0.7761**

Interprétation :

- L'hypothèse nulle du test de Bartlett est que les variances des groupes (ici, pour chaque condition météorologique) sont égales.
- La p-value (0.7761) est bien supérieure à un seuil commun de 0.05, ce qui signifie qu'il n'y a pas de preuve suffisante pour rejeter l'hypothèse nulle.
- Conclusion : Il n'y a pas de différence significative dans les variances entre les groupes (les variances sont homogènes).

2. Test de Kruskal-Wallis : Test des différences entre groupes (non paramétrique)

Le test de Kruskal-Wallis est un test non paramétrique utilisé pour comparer les médianes de plusieurs groupes indépendants. Contrairement au test de Bartlett, il ne nécessite pas que les données soient normalement distribuées.

- **Résultat du test de Kruskal-Wallis** :
  - **Chi-squared = 36.768**
  - **df = 4**
  - **p-value = 2.011e-07**

Interprétation :

- L'hypothèse nulle du test de Kruskal-Wallis est que les médianes des groupes (ici, pour chaque condition météorologique) sont égales.
- La p-value (2.011e-07) est extrêmement faible, bien inférieure au seuil de 0.05.

5) Faire un test robuste par bootsrap (rééchantillonage)

```
> PermTest(reg.aov, B=100)

Monte-Carlo test

Call:
PermTest.lm(obj = reg.aov, B = 100)

Based on 100 replicates
Simulated p-value:
                 p.value
henoc$Weather          0
> PermTest(reg.aov, B=1000)

Monte-Carlo test

Call:
PermTest.lm(obj = reg.aov, B = 1000)

Based on 1000 replicates
Simulated p-value:
                 p.value
henoc$Weather          0
```

Tableau 8 Boostrap

Conclusion : la p-value < 0.05 donc on rejette $H0$, Au moins une des moyennes est significativement différente des autres. Il y a donc bien l'existence d'un effet de la météo sur le temps de livraison.

6) Tester la significativité du facteur : tester l'égalité des moyennes

```
Analysis of Variance Table

Response: henoc$Delivery_Time_min
              Df Sum Sq Mean Sq F value    Pr(>F)
henoc$Weather   4  19545  4886.2  10.457 2.704e-08 ***
Residuals     965 450930   467.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

7) Analyser les résidus

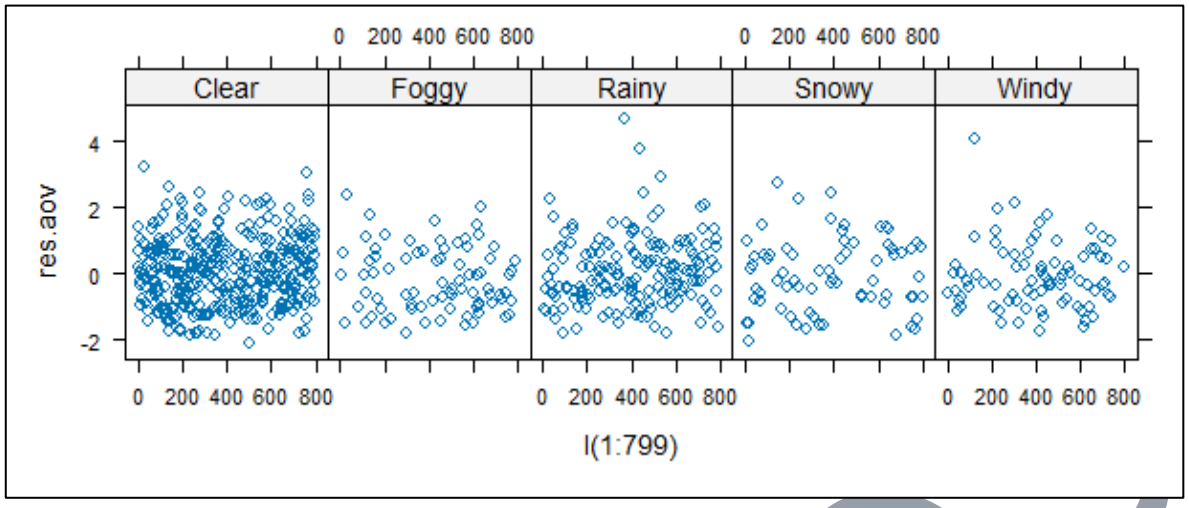


Figure 22Analyse des résidus

Conclusion : En théorie, 95% des résidus studentisés se trouvent dans l'intervalle [-2;2]. Ici, on constate que la grande majorité des résidus se trouvent dans cet intervalle.

```
> sum(as.numeric(abs(res.aov)<=2))/nrow(henoc)*100
[1] 93.4
```

8) Interpréter les coefficients

```
> anova(reg.aov)
Analysis of Variance Table

Response: henoc$Delivery_Time_min
               Df Sum Sq Mean Sq F value    Pr(>F)
henoc$Weather   4  19545  4886.2  10.457 2.704e-08 ***
Residuals     965 450930   467.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

9) Modèle ANOVA

Statistique	C(Weather)	Residual
sum_sq	19544.763158	450930.333750
df	4.000000	965.000000
F	10.456547	NaN
PR(>F)	0.000000	NaN

Statistique	Kruskal-Wallis
Statistique	37.0479
p-value	0.0000

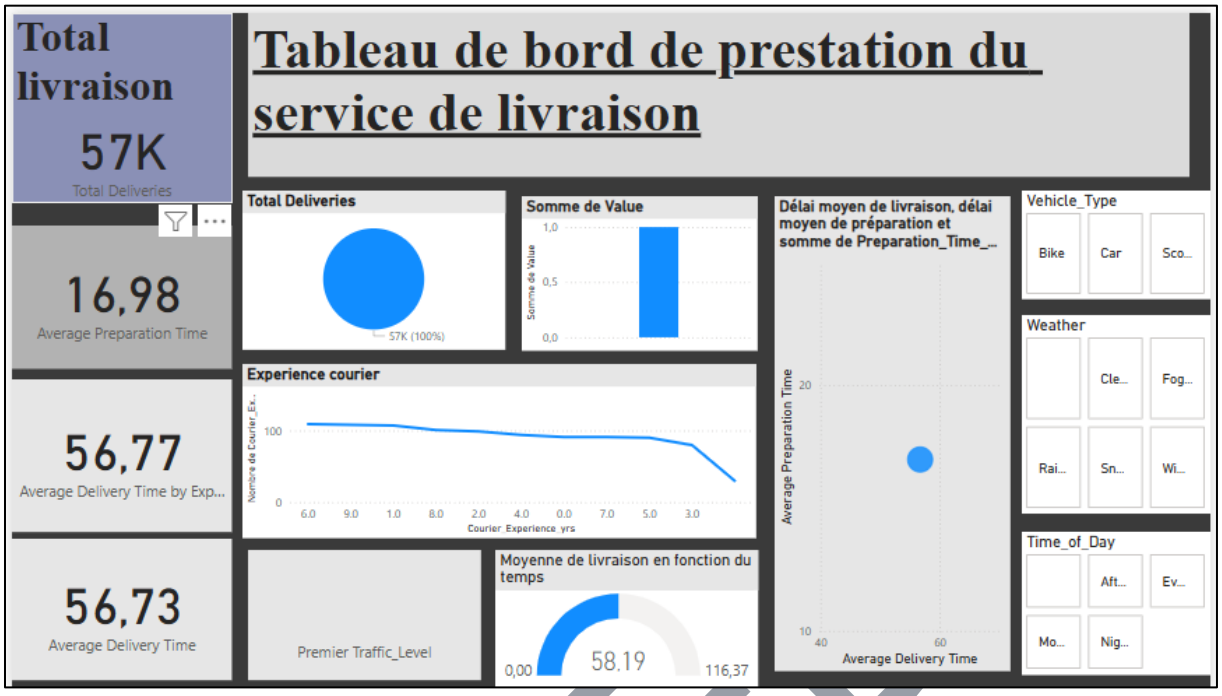
Statistique : 37.0479

p-valeur : 0.0000 (inférieure à 0.05)

Conclusion : La p-valeur très faible indique que les délais de livraison ne sont pas distribués de manière identique entre les groupes de conditions météorologiques.

Cela valide l'hypothèse selon laquelle la météo influence les délais de livraison.

Tableau de bord



CONCLUSION

L'objectif de cette analyse était de prédire les délais de livraison des commandes de nourriture en fonction de plusieurs facteurs influents tels que la distance, les conditions météorologiques, le trafic, l'heure de la journée et les caractéristiques des livreurs. Cette problématique est essentielle pour optimiser la logistique des livraisons et améliorer l'expérience client.

L'analyse a permis de démontrer que plusieurs variables, telles que la distance, le type de véhicule et l'expérience du livreur, ont une influence significative sur le délai de livraison. Grâce à la régression linéaire multiple, des relations quantitatives ont été identifiées entre ces facteurs et le délai total de livraison. De plus, l'ANOVA a révélé que les conditions météorologiques ont un impact notable sur le temps de livraison, avec des différences marquées selon les types de climat, ce qui confirme l'importance de prendre en compte ces éléments dans la gestion des livraisons.

Pour améliorer la gestion des délais de livraison, il est recommandé d'optimiser les itinéraires en fonction des prévisions de trafic et des conditions météorologiques, ainsi que d'ajuster les horaires de livraison selon les périodes de la journée pour tenir compte des variations du trafic. De plus, il serait bénéfique d'investir dans des formations pour les livreurs, notamment en gestion du temps et en adaptabilité face aux conditions climatiques diverses. Il pourrait également être judicieux de diversifier les types de véhicules utilisés, en fonction des conditions spécifiques rencontrées lors des livraisons.

Cependant, il convient de noter que les résultats de cette analyse sont limités par la disponibilité des données et leur représentativité. Par exemple, la variable "Météo" a été simplifiée en catégories génériques (clair, pluvieux, neigeux, etc.), ce qui ne permet pas de saisir toute la complexité des effets réels de la météo. En outre, le modèle ne prend pas en compte certains facteurs externes tels que des événements imprévus ou des comportements individuels des livreurs, qui pourraient également influencer les délais de livraison.

Pour aller plus loin et améliorer la précision des prédictions, plusieurs pistes peuvent être explorées. Il serait pertinent d'introduire des variables supplémentaires, comme la densité urbaine, la gestion dynamique du trafic en temps réel, ou encore d'intégrer les horaires de livraison en fonction des fluctuations de la demande. Une analyse plus détaillée des conditions météorologiques, utilisant des données précises telles que la température, l'humidité ou la vitesse du vent, pourrait offrir des insights supplémentaires. Enfin, l'adoption de modèles plus complexes, comme les réseaux neuronaux ou les arbres décisionnels, permettrait de mieux comprendre les interactions non linéaires et d'améliorer la modélisation des délais de livraison.

ANNEXE

Modèles et outils statistiques

Scikit-learn: scikit-learn: machine learning in Python — scikit-learn 1.6.1 documentation

Tableau Public : Superstore Performance Dashboard || Dynamic KPI Color with Dark-Light Mode | Tableau Public

National Weather Service : National Weather Service

1. Régression Linéaire Multiple

Scikit-learn Documentation : Régression linéaire

- Régression Linéaire dans scikit-learn : Ce lien vous fournit la documentation officielle de la bibliothèque scikit-learn, où vous pouvez trouver des informations détaillées sur l'implémentation de la régression linéaire multiple en Python.

- **Kaggle : Tutoriel sur la Régression Linéaire**
  - Introduction à la Régression Linéaire sur Kaggle : Un tutoriel interactif pour débuter avec la régression linéaire, avec des exemples pratiques sur Kaggle.

- **Statistical Methods for Regression Analysis (Coursera)**
  - Coursera - Introduction à la Régression Linéaire : Cours complet de Coursera pour comprendre les méthodes statistiques derrière la régression linéaire et leur application pratique.

2. ANOVA (Analyse de la Variance)

- **Statsmodels Documentation : ANOVA**
  - ANOVA dans Statsmodels : La documentation officielle de statsmodels pour effectuer une analyse de la variance (ANOVA) avec des données catégorielles en Python.

- **Khan Academy : Comprendre l'ANOVA**
  - Vidéo d'Introduction à l'ANOVA sur Khan Academy : Une explication simple de l'ANOVA, idéale pour les débutants souhaitant comprendre les concepts théoriques.

- **TutorialsPoint : ANOVA avec Python**
  - Tutoriel sur l'ANOVA avec Python : Ce lien offre un tutoriel détaillé pour effectuer une ANOVA en Python, en utilisant des bibliothèques telles que scipy et statsmodels.

3. Ressources générales pour les projets de régression linéaire et ANOVA

- **Real Python : Introduction à la Régression Linéaire avec Python**
  - Real Python - Régression Linéaire : Un tutoriel détaillé sur la régression linéaire avec Python, couvrant à la fois les aspects théoriques et pratiques.

- **Medium : Tutoriel sur ANOVA avec Python**
  - Medium - Tutoriel ANOVA en Python : Un article pratique sur la façon de réaliser un test ANOVA en Python avec statsmodels et scipy.

Source du code R et python

```
##############################################
                "REGRESSION LINEAIRE"
##############################################
"1ere étape : Importer les données"
#######################################
library(tidymodels)
library(stringr)
library(readr)
henoc <- read_csv("C:/Users/yoboh/OneDrive/Bureau/INSSEDS/MINI PROJET/PROJETS/ECONOMETRIE/LIVRAISON/Data BASE/delai_livraison.csv")
print(henoc)
str(henoc)
summary(henoc)
# Supposons que votre dataframe s'appelle df
henoc$Order_ID <- as.character(henoc$Order_ID)
summary(henoc)

"2eme étape : Représenter les variables"
#########################################
#supprimer la colonne order id
henoc <- henoc[, !names(henoc) %in% c("Order_ID")]

#selection des colonne numerique
henoc_numeric <- henoc[, sapply(henoc, is.numeric)]  # Sélectionne seulement les colonnes numériques
cor_matrix <- cor(henoc_numeric, use = "complete.obs")
cor(henoc_numeric,use="complete.obs")

# convertir les colonne categorielle en num
# Si vous avez des colonnes catégorielles comme 'Weather', 'Traffic_Level', etc.
henoc$Weather <- as.numeric(factor(henoc$Weather))
henoc$Traffic_Level <- as.numeric(factor(henoc$Traffic_Level))
henoc$Time_of_Day <- as.numeric(factor(henoc$Time_of_Day))
henoc$Vehicle_Type <- as.numeric(factor(henoc$Vehicle_Type))
cor(henoc,use="complete.obs")
```

```
# b) Visualiser les corrélations deux a deux
pairs(henoc)
cor(henoc,use="complete.obs")
cor(henoc, use="pairwise.complete")

##c) Visualiser les corrélations deux a deux
library(corrplot)
cor<- cor(henoc, use="pairwise.complete")
corrplot(cor)

#d) Visualiser les corrélations deux a deux (meilleur choix )
library(ggplot2)
ggpairs(henoc)

#e) ACP avec l'endogène en illustratif
library(FactoMineR)
res.pca<-PCA(henoc,quanti.sup=1)
res.acm <- MCA(henoc, graph = TRUE)

# ACM sous shiny
library(Factoshiny)
res <- MCAshiny(henoc)
res.MCA <- MCA(henoc,graph=FALSE)

"3eme étape : Estimer les paramètres "
# +-------------------------------+#
#####################################

# Faire la régression
regM <- lm(Delivery_Time_min ~ Distance_km + Preparation_Time_min +
Courier_Experience_yrs + Weather + Traffic_Level + Time_of_Day + Vehicle_Type, data
= henoc)
regM$coefficients

# Voir l'intervalle de confiance des coefficients estimés
```

```
confint(regM)

summary(regM)


#Convertir les variables catégorielles en variables factorielles

henoc$Weather <- factor(henoc$Weather)

henoc$Traffic_Level <- factor(henoc$Traffic_Level)

henoc$Time_of_Day <- factor(henoc$Time_of_Day)

henoc$Vehicle_Type <- factor(henoc$Vehicle_Type)


# Régression linéaire multiple

model <- lm(Delivery_Time_min ~ Distance_km + Preparation_Time_min +
Courier_Experience_yrs + Weather + Traffic_Level + Time_of_Day + Vehicle_Type, data
= henoc)


# Résumé du modèle

summary(model)

confint(model)

model$coefficients

##########################################################################
#######################

            "MODELE ET SELECTION AUTOMATIQUE DES EXPLICATIVES"

##########################################################################
#######################

# Traitement de données manquantes

henoc <- na.omit(henoc)


# modèle stepwise backward

modele_complet <- lm(Delivery_Time_min ~ . , data = henoc)

modele_2 <- step(modele_complet, direction="backward")

print(modele_complet)


install.packages("MASS")

library(MASS)

modele_2 <- stepAIC(modele_complet, direction = "backward")

summary(modele_2)
```

```
"# 4eme étape : Choix de variables "

# +--------------------------+#

#####################################

library(leaps)

choix<-regsubsets(henoc$Delivery_Time_min ~ henoc$Distance_km + henoc$Weather
+ henoc$Traffic_Level +

            henoc$Preparation_Time_min +henoc$Courier_Experience_yrs
,data=henoc,nbest=2,nvmax=5)

plot(choix,scale="bic")


#Faisons à nouveau la régression sur les variables retenues


reg.fin<-lm(henoc$Delivery_Time_min ~ henoc$Distance_km +
henoc$Preparation_Time_min + henoc$Traffic_Level + henoc$Weather)

reg.fin$coefficients

confint(reg.fin)

summary(reg.fin)


"# 5eme étape : Analyser les résidus "

# +----------------------------+#

res.m<-rstudent(reg.fin)

plot(res.m,pch=15,cex=.5,ylab="Residus",ylim=c(-3,3))

abline(h=c(-2,0,2),lty=c(2,1,2), color = "red")


res.m<-rstudent(reg.fin)

sum(as.numeric(abs(res.m)<=2))/nrow(df)*100


install.packages("olsrr")

library(olsrr)

ols_vif_tol(reg.fin)


"6eme étape : Tester la validité du modèle"

# +-------------------------------------+#

# Test de normalité des résidus (Shapiro-Wilk)

shapiro.test(residuals)
```

```
# Q-Q plot des résidus
qqnorm(residus)
qqline(residus, col = "red")

residus<-residuals(reg.fin)
res.normalise<-rstudent(reg.fin)
val.estimees<-fitted.values(reg.fin)

##1) Test de linéarité du modèle
library(lmtest)
raintest(reg.fin)

#2) Test si de  Cov(X,ε)=0
plot(reg.fin,1)

###### homoscédasticité
0 : il y a homoscédasticité
1 : il y a hétéroscédasticité
install.packages("lmtest")
library(lmtest)
bptest(reg.fin)

#4) Test d'autocorrelationdes erreurs Cov(εt,εs) = 0
res.m<-rstudent(reg.fin)
plot(res.m, pch=15, cex=0.5, ylab="Residus", ylim=c(-3, 3), type="b")
acf(residus,plot = FALSE)
acf(residus,plot = TRUE)

####TEST DE DURBAN - WATSON
install.packages("lmtest")
library(lmtest)
dwtest(reg.fin)
```

#5) Test de $E(\varepsilon t)=0$

```
mean(residus)
```

##6) $\varepsilon t \sim N(0, \sigma\varepsilon^2)$ normalité des erreurs

TEST DE SHAPIRO WILK

 # $H0$: La distribution suit la loi normale

 # $H1$: La distribution ne suit pas la loi normale

```
shapiro.test(residus)
```

##Test de Chow: stabilité ou rupture de structure

```
install.packages("strucchange")
library(strucchange)
sctest(henoc$Delivery_Time_min ~ henoc$Distance_km + henoc$Preparation_Time_min
+ henoc$Traffic_Level + henoc$Weather, type = "Chow")
```

"7eme étape : Prédire une nouvelle valeur"

```
# +-----------------------------------+#
 ##########################################
 henoc1<-
data.frame(henoc$Delivery_Time_min,henoc$Distance_km,henoc$Preparation_Time_mi
n,henoc$Traffic_Level,henoc$Weather, type = "Chow")
 reglm<-lm(henoc.Delivery_Time_min ~ henoc.Distance_km +
henoc.Preparation_Time_min + henoc.Traffic_Level + henoc.Weather, data=henoc1)
 summary(reglm)

Xnew<-matrix(c(45,1,0),nrow=1)
colnames(Xnew)<-
c("henoc.Distance_km","henoc.Preparation_Time_min","henoc.Traffic_Level2")
Xnew<-as.data.frame(Xnew)
Xnew

predict(reglm, Xnew, interval="pred")

# Étape 1 : Convertir Traffic_Level et Weather en facteurs (si ce n'est pas déjà fait)
```

```
henoc$Traffic_Level <- as.factor(henoc$Traffic_Level)

henoc$Weather <- as.factor(henoc$Weather)


# Vérifier les niveaux des facteurs

levels(henoc$Traffic_Level)  # Voir les niveaux de Traffic_Level

levels(henoc$Weather)        # Voir les niveaux de Weather


# Étape 2 : Créer le modèle de régression linéaire

henoc1 <- data.frame(henoc$Delivery_Time_min, henoc$Distance_km,
henoc$Preparation_Time_min, henoc$Traffic_Level, henoc$Weather, type = "Chow")

reglm <- lm(henoc.Delivery_Time_min ~ henoc.Distance_km +
henoc.Preparation_Time_min + henoc.Traffic_Level + henoc.Weather, data = henoc1)


# Afficher le résumé du modèle

summary(reglm)


# Étape 3 : Créer un nouveau jeu de données pour la prédiction

# Assurez-vous que les variables 'Traffic_Level' et 'Weather' sont des facteurs avec les
bons niveaux


Xnew <- data.frame(

  henoc.Distance_km = 45,  # Valeur pour Distance_km

  henoc.Preparation_Time_min = 1,  # Valeur pour Preparation_Time_min

  # Utiliser les bons niveaux pour Traffic_Level et Weather

  henoc.Traffic_Level = factor("Medium", levels = levels(henoc$Traffic_Level)),  #
Exemple de niveau dans Traffic_Level

  henoc.Weather = factor("Clear", levels = levels(henoc$Weather))  # Exemple de
niveau dans Weather

)


# Vérifier la structure de Xnew

str(Xnew)


# Étape 4 : Faire la prédiction avec l'intervalle de prédiction

predictions <- predict(reglm, Xnew, interval = "prediction")


# Afficher les prédictions
```

```
print(predictions)

# Exemple de données dans Xnew

Xnew <- matrix(c(45, 1, 0, "cloudy"), nrow = 1)  # Remplacer "sunny" par un niveau valide comme "cloudy"

colnames(Xnew) <- c("henoc.Distance_km", "henoc.Preparation_Time_min", "henoc.Traffic_Level", "henoc.Weather")

Xnew <- as.data.frame(Xnew)


# Convertir les variables Traffic_Level et Weather en facteurs

Xnew$henoc.Traffic_Level <- factor(Xnew$henoc.Traffic_Level, levels = c("low", "medium", "high"))

Xnew$henoc.Weather <- factor(Xnew$henoc.Weather, levels = c("cloudy", "rainy", "clear"))


# Effectuer la prédiction

predict(reglm, Xnew, interval = "pred")




"9eme étape : Modèle avec interaction"

########################################


# Ajouter une interaction entre la distance et le type de véhicule

model_interaction <- lm(Delivery_Time_min ~ Distance_km * Vehicle_Type + Preparation_Time_min + Courier_Experience_yrs + Weather + Traffic_Level + Time_of_Day, data = henoc)

# Résumé du modèle avec interaction

summary(model_interaction)



################################################

                "PARTIE III : ANOVA"

################################################

henoc <- read_csv("C:/Users/yoboh/OneDrive/Bureau/INSSEDS/MINI PROJET/PROJETS/ECONOMETRIE/LIVRAISON/Data BASE/delai_livraison.csv")



## 2eme étape : estimer les statistiques de base (mean, quantile, sd) par ss pop
```

```
tapply(henoc$Delivery_Time_min,henoc$Weather,mean,na.rm=TRUE)
tapply(henoc$Delivery_Time_min,henoc$Weather,sd,na.rm=TRUE)
tapply(henoc$Delivery_Time_min,henoc$Weather,quantile,na.rm=TRUE)

## 3eme étapetester la normalité des données dans chaque sous population
library(car)
library(RVAideMemoire)
byf.qqnorm(henoc$Delivery_Time_min~henoc$Weather)
par(mar = c(4, 4, 2, 2))
byf.qqnorm(henoc$Delivery_Time_min ~ henoc$Weather)

library(RVAideMemoire)
byf.shapiro(henoc$Delivery_Time_min~henoc$Weather)

## 4eme étape : tester l'égalité des variances
bartlett.test(henoc$Delivery_Time_min~henoc$Weather)
kruskal.test(henoc$Delivery_Time_min~henoc$Weather)

## 5eme étape : faire un test robuste par bootsrap (rééchantillonage)
install.packages("pgirmess")
library(pgirmess)
reg.aov<-lm(henoc$Delivery_Time_min~henoc$Weather)
PermTest(reg.aov, B=100)
PermTest(reg.aov, B=1000)
PermTest(reg.aov, B=500)


## 5eme étape : tester la significativité du facteur: tester l'égalité des moyennes
reg.aov<-lm(henoc$Delivery_Time_min~henoc$Weather)
anova(reg.aov)

install.packages("coin")
library(coin)
pairwise.perm.t.test(henoc$Delivery_Time_min, henoc$Weather)
```

6eme étape : Analyser les résidus

library(lattice)

res.aov<-rstudent(reg.aov)

xyplot(res.aov~I(1:799)|henoc$Weather)

res.aov<-rstudent(reg.aov)

sum(as.numeric(abs(res.aov)<=2))/nrow(henoc)*100

7eme étape : Interpréter les coefficients

reg.aov<-lm(henoc$Delivery_Time_min~henoc$Weather)

anova(reg.aov)

##1 1) Calcul des coefficients avec $\alpha 1 = 0$ option par défaut sur R : smp$agriculteur étant la première modalité par ordre alphabétique, c'est e

eg.aov<-lm(henoc$Delivery_Time_min~henoc$Weather)

summary(reg.aov)