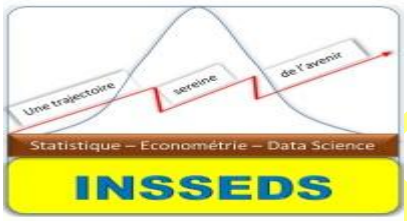
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

REPUBLIQUE DE CÔTE D'IVOIRE

Union – Discipline – Travail



INSTITUT SUPERIEUR DE STATISTIQUE, D'ECONOMETRIE ET DE DATASCIENCE

MASTER 2

STATISTIQUE ECONOMETRIE DATA SCIENCE

MINI PROJET ECONOMETRIE DES VARIABLES QUANTITATIVES

ETUDE DES VARIABLES INFLUENCANT LE DELAI DE LIVRAISONA L'AIDE DE METHODES ECONOMETRIQUE

MODELE REGRESSION MULTIPLE ET MODELE ANOVA-1

ANNEE UNIVERSITAIRE : 2024-2025

ETUDIANT :

N'DRI ONESIME

ENSEIGNANT :

AKPOSSO DIDIER MARTIAL

AVANT-PROPOS

L'association du théorique au pratique, des connaissances aux compétences et des savoir-faire aux savoirs est la principale tendance récente dans le secteur technique. Dans ce contexte, l'INSSEDS (Institut Supérieur de la Statistique, d'Econométrie et de la Data Science), dans sa formation en master professionnel en statistique, économie et science des données, impose que les divers crédits soient validés en effectuant un mini-projet à la fin de chaque module. Le projet est donc structuré et supervisé de cette manière, visant principalement à faire de chaque élève un participant dynamique, engagé et libre dans la vie active.

Ce document est un rapport de projet Muni axé sur l'économétrie des variables quantitatives. Il se divise principalement en trois parties : Prétraitements des données, Analyse univariée, Analyse bivariée, Estimation du Délai moyen de livraison, Test de conformité à la moyenne du Délai moyen de livraison, Test de liaison de la variables cible avec les variables sélectionnées, Modélisation Régression linéaire multiple, Modélisation ANOVA-1.

En règle générale, toutes les analyses et conclusions présentées dans ce rapport relèvent de la responsabilité de l'auteur, qui ne sollicite ni autrui ni l'INSSEDS (Institut Supérieur de Statistique d'Econométrie et de Data Science).

Table des matières

INTRODUCTION

Contexte et justification de l'étude

Avec l'augmentation du commerce en ligne et la demande croissante de services de livraison rapide, il est crucial de pouvoir estimer avec précision les délais de livraison pour mieux planifier et gérer les opérations de livraison. Une estimation précise des délais permet aux entreprises de réduire les coûts opérationnels, d'améliorer la satisfaction des clients et de maintenir une compétitivité sur le marché.

Problématique

Comment les divers facteurs tels que la distance, la météo, les conditions de circulation, l'heure de la journée, le type de véhicule, le temps de préparation de la commande et l'expérience du coursier influencent-ils les délais de livraison ? Comprendre ces influences permettrait de développer des stratégies pour optimiser les processus de livraison et réduire les délais.

Principaux résultats attendus

Développer des modèles prédictifs précis pour estimer les délais de livraison et identifier les facteurs les plus influents. Ces modèles permettront aux entreprises de mieux planifier leurs opérations, d'anticiper les retards potentiels et de prendre des mesures proactives pour améliorer l'efficacité de la livraison.

Méthodologie

La méthodologie proposée comprend : - Le prétraitement des données - L'analyse univariée et bivariée - L'estimation de la moyenne de la variable Delivery_time_min - Le test de conformité à un standard de la moyenne - Le test de liaison - La modélisation de régression multiple - La modélisation ANOVA

Description du jeu de données : dictionnaire des données

- **Order_ID : Identifiant unique pour chaque commande.**
- **Distance_km : La distance de livraison en kilomètres.**
- **Météo : Conditions météorologiques pendant la livraison, y compris clair, pluvieux, neigeux, brumeux et venteux.**
- **Traffic_Level : conditions de trafic classées comme faibles, moyennes ou élevées.**
- **Time_of_Day : L'heure à laquelle la livraison a eu lieu, classée comme Matin, Après-midi, Soir ou Nuit.**
- **Vehicle_Type : Type de véhicule utilisé pour la livraison, y compris vélo, scooter et voiture.**
- **Preparation_Time_min : Le temps nécessaire à la préparation de la commande, mesuré en minutes.**
- **Courier_Experience_yrs : Expérience du coursier en années.**
- **Delivery_Time_min : Le délai de livraison total en minutes (variable cible).**

I- Prétraitements des données

Avant l'entame de la deuxième partie du Prétraitement des données, il est important de signifier que quelques modifications ont été appliqués sur le jeu de données telles que la modification des types de données, le renommage des modalités de quelques variables, le traitement des doublons. Tout cela s'est fait dans le logiciel Excel-Power-Query.

Visualisation des données

```
tibble [1,000 × 9] (S3: tbl_df/tbl/data.frame)
Order_ID               : chr [1:1000] "522" "738" "741" "661" ...
Delivery_Time_min      : num [1:1000] 43 84 59 37 68 57 49 46 35 73 ...
Distance_km            : num [1:1000] 7.93 16.42 9.52 7.44 19.03 ...
Weather                : chr [1:1000] "Venteux" "Clair" "Brumeux" "Plu-
vieux" ...
Traffic_Level          : chr [1:1000] "Faible" "Moyenne" "Faible"
"Moyenne" ...
Time_of_Day            : chr [1:1000] "Apres-midi" "Soir" "Nuit"
"Apres-midi" ...
Vehicle_Type           : chr [1:1000] "Scooter" "Velo" "Scooter" "Scoo-
ter" ...
Preparation_Time_min   : num [1:1000] 12 20 28 5 16 8 12 5 20 29 ...
Courier_Experience_yrs: num [1:1000] 1 2 1 1 5 9 1 6 6 1 ...
```

Ce jeu de données comprend 1000 observations et 9 variables que sont Order_ID,Distance_km,Weather,Traffic_Level,Time_of_Day,Vehicle_Type, Preparation_Time_min, Courier_Experience_yrs et Delivery_Time_min.

Traitement des valeurs manquantes

D'abord nous allons changer les variables de chaines de characteres en factor.

```
A tibble: 1,000 × 9
   Order_ID Delivery_Time_min Distance_km Weather  Traffic_Level Time_of_Day
   <chr>                <dbl>       <dbl> <fct>    <fct>         <fct>
 1 522                     43        7.93 Venteux  Faible        Apres-midi
 2 738                     84       16.4  Clair    Moyenne       Soir
 3 741                     59        9.52 Brumeux  Faible        Nuit
 4 661                     37        7.44 Pluvieux Moyenne       Apres-midi
 5 412                     68       19.0  Clair    Faible        Matin
 6 679                     57       19.4  Clair    Faible        Soir
 7 627                     49        9.52 Clair    Faible        <NA>
 8 514                     46       17.4  Clair    Moyenne       Soir
 9 860                     35        1.78 Neigeux  Faible        Soir
10 137                     73       10.6  Brumeux  Faible        Soir
ℹ 990 more rows
ℹ 3 more variables: Vehicle_Type <fct>, Preparation_Time_min <dbl>,
  Courier_Experience_yrs <dbl>
```

Résumé des variables avant traitement des valeurs manquantes :

```
   Order_ID          Delivery_Time_min  Distance_km         Weather
 Length:1000        Min.   :  8.00    Min.   : 0.590   Brumeux :103
 Class :character   1st Qu.: 41.00    1st Qu.: 5.105   Clair   :470
 Mode  :character   Median : 55.50    Median :10.190   Neigeux : 97
                    Mean   : 56.73    Mean   :10.060   Pluvieux:204
                    3rd Qu.: 71.00    3rd Qu.:15.018   Venteux : 96
                    Max.   :153.00    Max.   :19.990   NA's    : 30

 Traffic_Level     Time_of_Day   Vehicle_Type Preparation_Time_min
 Elevee :197   Apres-midi:284   Scooter:302   Min.   : 5.00
 Faible :383   Matin      :308   Velo   :503   1st Qu.:11.00
 Moyenne:390   Nuit       : 85   Voiture:195   Median :17.00
 NA's   : 30   Soir       :293                 Mean   :16.98
               NA's       : 30                 3rd Qu.:23.00
                                               Max.   :29.00

 Courier_Experience_yrs
 Min.   :0.000
 1st Qu.:2.000
 Median :5.000
 Mean   :4.579
 3rd Qu.:7.000
 Max.   :9.000
 NA's   :30
```

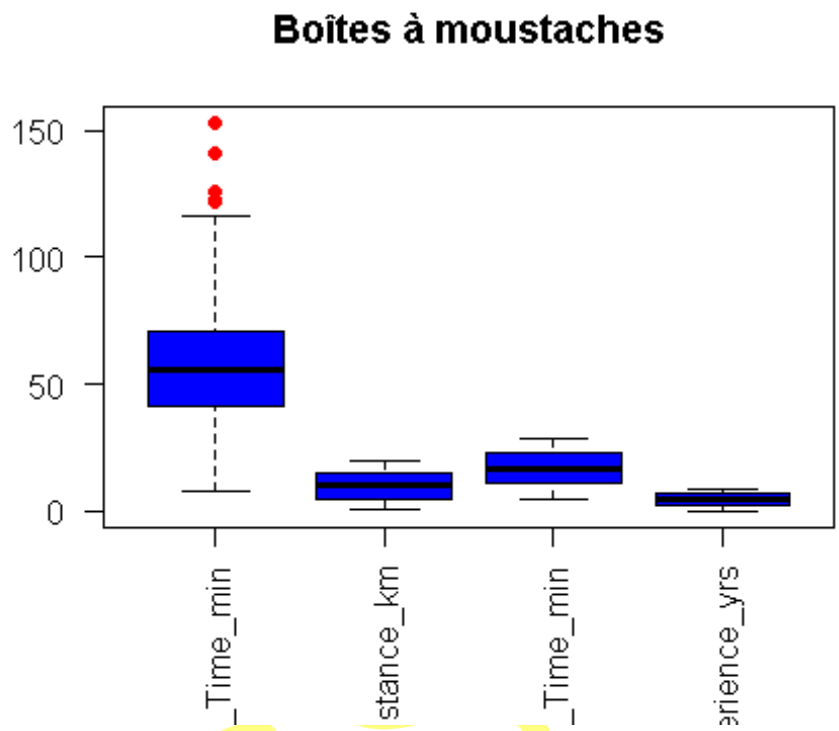
On peut s'apercevoir qu'il y'a 30 valeurs manquantes dans les variables "Weather", "Traffic_Level", "Time_of_Day" et "Courier_Experience_yrs"

Résumé des variables après traitement des valeurs manquantes :

```
   Order_ID          Delivery_Time_min  Distance_km         Weather
 Length:1000        Min.   :  8.00    Min.   : 0.590   Brumeux :103
 Class :character   1st Qu.: 41.00    1st Qu.: 5.105   Clair   :489
 Mode  :character   Median : 55.50    Median :10.190   Neigeux : 99
                    Mean   : 56.73    Mean   :10.060   Pluvieux:212
                    3rd Qu.: 71.00    3rd Qu.:15.018   Venteux : 97
                    Max.   :153.00    Max.   :19.990
 Traffic_Level     Time_of_Day   Vehicle_Type Preparation_Time_min
 Elevee :201   Apres-midi:296   Scooter:302   Min.   : 5.00
 Faible :399   Matin      :319   Velo   :503   1st Qu.:11.00
 Moyenne:400   Nuit       : 87   Voiture:195   Median :17.00
               Soir       :298                 Mean   :16.98
                                               3rd Qu.:23.00
                                               Max.   :29.00

 Courier_Experience_yrs
 Min.   :0.000
 1st Qu.:2.000
 Median :5.000
 Mean   :4.603
 3rd Qu.:7.000
 Max.   :9.000
```

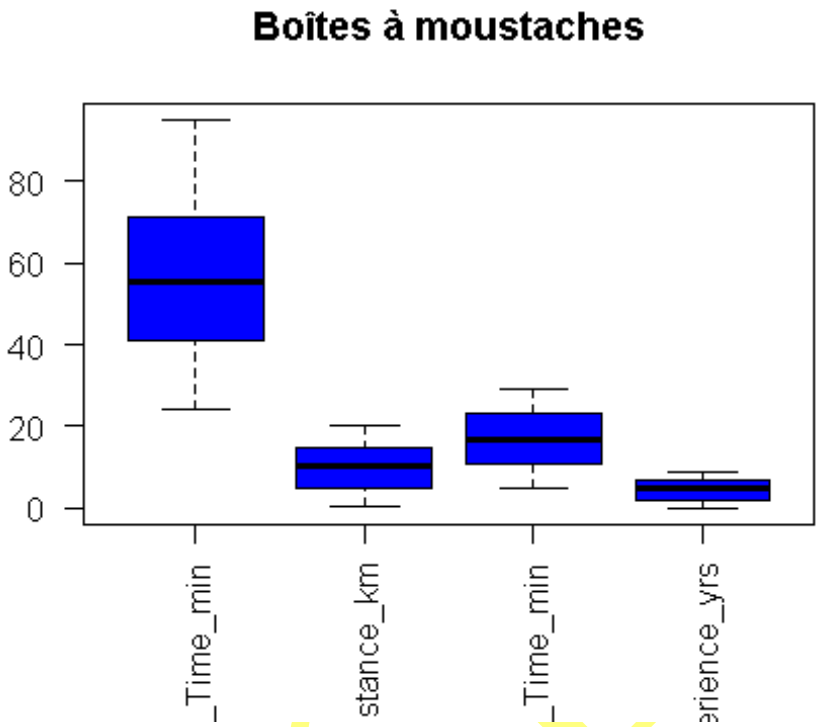
Traitement des valeurs abberantes

- **visualisation des valeurs abberantes**



On remarque que la variable Delivery_Time_min comporte des valeurs manquantes car 75% des valeurs gravites autours de 71 minutes et le temps maximal est de 153 minutes. Ainsi, il convient de les traiter.

- **Visualisation après traitement**



Les valeurs aberrantes ainsi traiter, nous pouvons passer aux différentes analyses du jeu de données.

II- ANALYSE UNIVARIEE

II-1 ANALYSE DES VARIABLES QUANTITATIVES

Variable "Delivery_Time_min"

- **Tableau**

	Effectifs	Eff_Cum_crois	Eff_Cum_decrois	Frequence	Freq_Cum_crois	Freq_Cum_decrois
24	52	52	1000	0.052	0.052	1.000
25	7	59	948	0.007	0.059	0.948
26	6	65	943	0.006	0.065	0.943
27	13	78	940	0.013	0.078	0.940
28	17	95	932	0.017	0.095	0.932

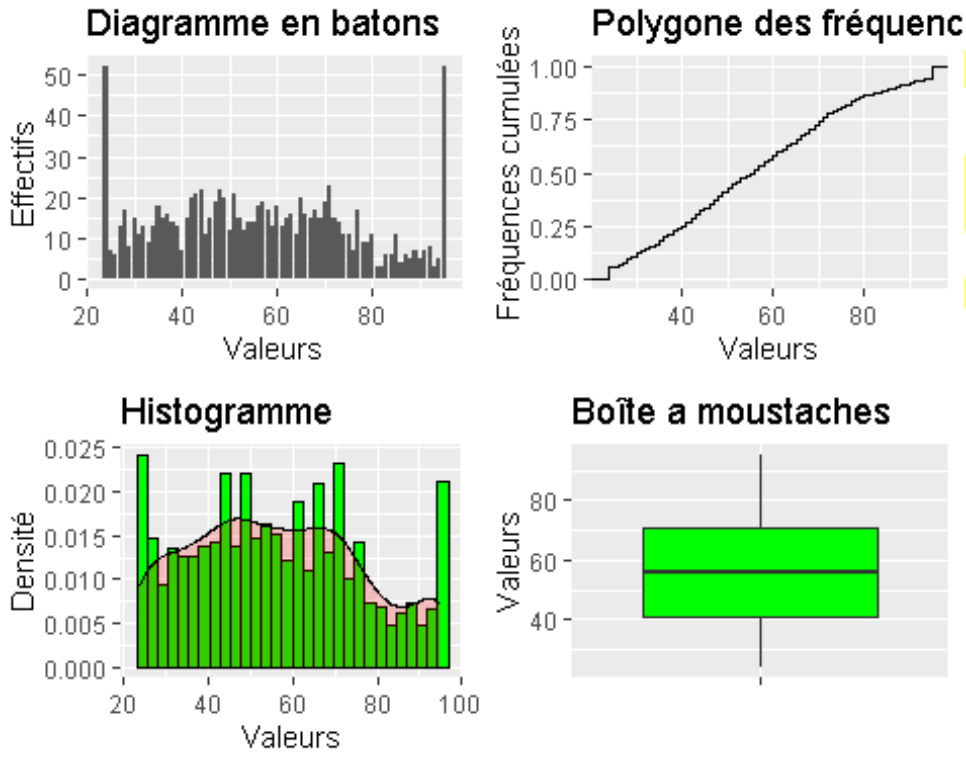
95 livraisons font au plus 28 min

932 Livraisons font au moins 28 min

9,5% des livraisons font au plus 28 min

93,2% des livraisons font au moins 28min

- **Graphique**



Le diagramme en bâtons nous permet de dire que la plupart des temps de livraisons sont entre 24 min et 95 min

Le polygone des fréquences cumulées nous permet de dire que plus le temps de livraisons augmente plus la fréquence des livraisons augmente

Aux vues de l'histogramme, la variable "Delivery_Time_min" ne semble pas suivre une loi normale

Le boxplot nous permet de dire que 50% des temps de livraisons sont autour de 56 min

- **Resumés numériques**

```
Minimum
 [1] 24

 maximum
 [1] 95

 mode
 [1] 24 95

 médiane
 [1] 55.5

 moyenne
 [1] 56.334

 quantile
   0%  25%  50%  75% 100%
 24.0 41.0 55.5 71.0 95.0

 coefficient_variation
 [1] 35.75364

 variance
 [1] 405.6781

 ecart_type
 [1] 20.14145

 coefficient_assymetrie
 [1] 0.2183673

 interpretation_skewness
 [1] "distribution etalee a droite"

 coefficent_applatissement
 [1] 2.128246

 interpretation_kurtosis
 [1] "distribution platikurtique"
```

Variable “Distance_km”

- **Tableau**

	Effectifs	Eff_Cum_crois	Eff_Cum_decrois	Frequence	Freq_Cum_crois	Freq_Cum_decrois
0.59	1	1	1000	0.001	0.001	1.000
0.6	1	2	999	0.001	0.002	0.999
0.61	1	3	998	0.001	0.003	0.998
0.64	1	4	997	0.001	0.004	0.997
0.68	1	5	996	0.001	0.005	0.996

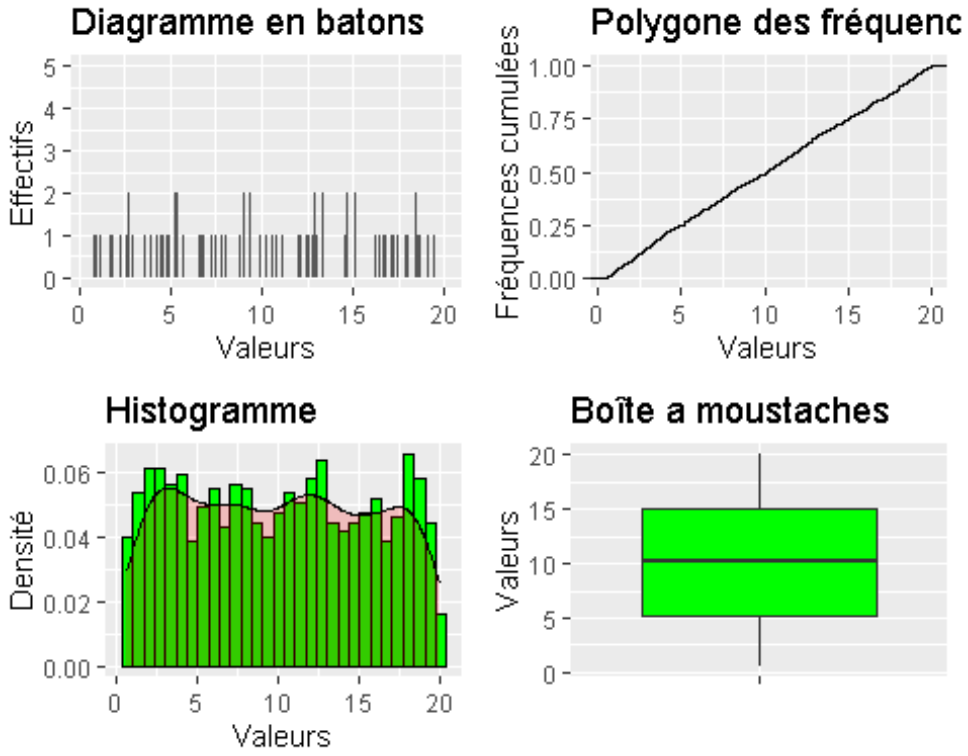
5 livraisons font au plus 0,68Km

996 livraisons font au moins 0,68Km

0,5% des livraisons font au plus 0,68Km

99,6% des livraisons font au moins 0,68Km

- **Graphique**



Le diagramme en bâtons nous permet de dire que la plupart des distances de livraisons font entre 0,68Km et 19Km

Le polygone des fréquences cumulées nous permet de dire que plus la distance de livraisons augmente plus la fréquence des livraisons augmente

Aux vues de l'histogramme, la variable "Distance_Km" ne semble pas suivre une loi normale

Le boxplot nous permet de dire que 50% des temps de livraisons sont autour de 10 Km

- **Résumé numérique**

```
minimum
 [1] 0.59

 maximum
 [1] 19.99

 mode
 [1] 3.149906

 mediane
 [1] 10.19

 moyenne
 [1] 10.05997

 quantile
     0%     25%     50%     75%    100%
  0.5900  5.1050 10.1900 15.0175 19.9900

 coefficient_variation
 [1] 56.62696

 variance
 [1] 32.45188

 ecart_type
 [1] 5.696656

 coefficient_assymetrie
 [1] 0.03878219

 interpretation_skewness
 [1] "distribution etalee a droite"

 coefficent_applatissement
 [1] 1.770896

 interpretation_kurtosis
[1] "distribution platikurtique"
```

II-2 ANALYSE DES VARIABLES QUALITATIVES

Variable “Météo”

- **Tableau**

	Effectif	Fréquence
Brumeux	103	0.103
Clair	489	0.489
Neigeux	99	0.099
Pluvieux	212	0.212
Venteux	97	0.097

103 livraisons ou 10,3% des livraisons se font dans un temps Brumeux

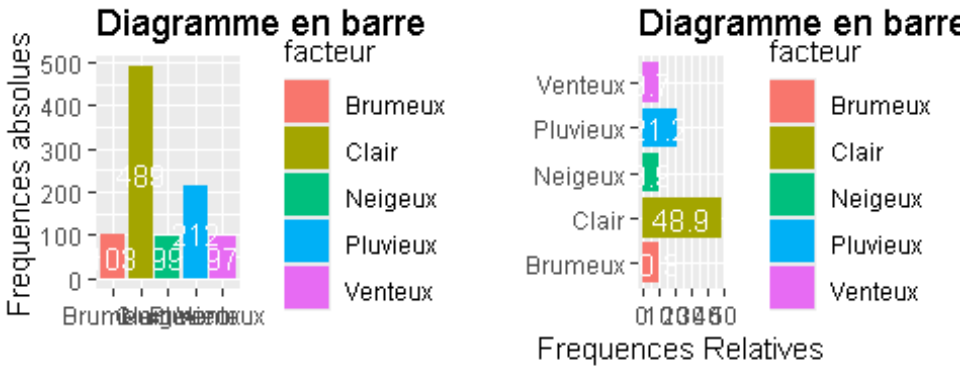
489 livraisons ou 48,9% des livraisons se font dans un temps Clair

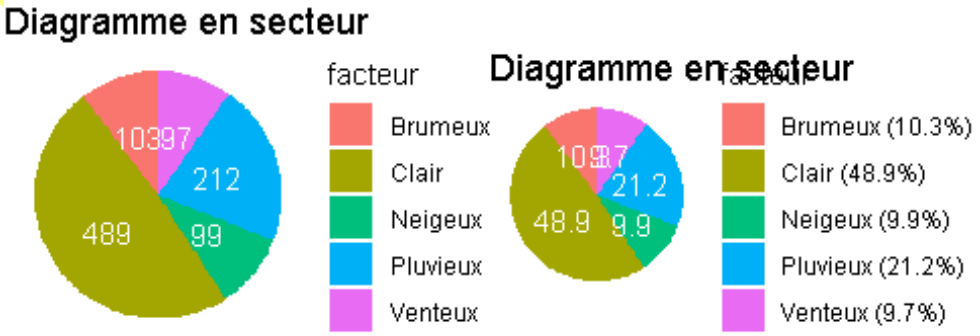
99 livraisons ou 9,9% des livraisons se font dans un temps Neigeux

212 livraisons ou 21,2% des livraisons se font dans un temps Pluvieux

97 livraisons ou 9,7% des livraisons se font dans un temps Venteux

- **Graphique**





Le plus grand nombre de livraisons se fait dans un temps clair et occupent 48,7% des livraisons totales. En revanche, le plus petit nombre de livraisons se fait dans un temps Neigeux et occupent 10% des livraisons totales.

Variables "Traffic Level"

- **Tableau**

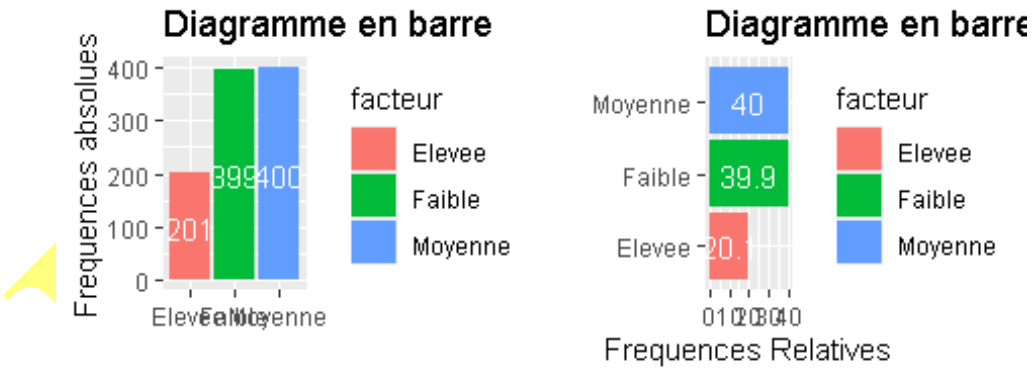
	Effectif	Frequence
Elevee	201	0.201
Faible	399	0.399
Moyenne	400	0.400

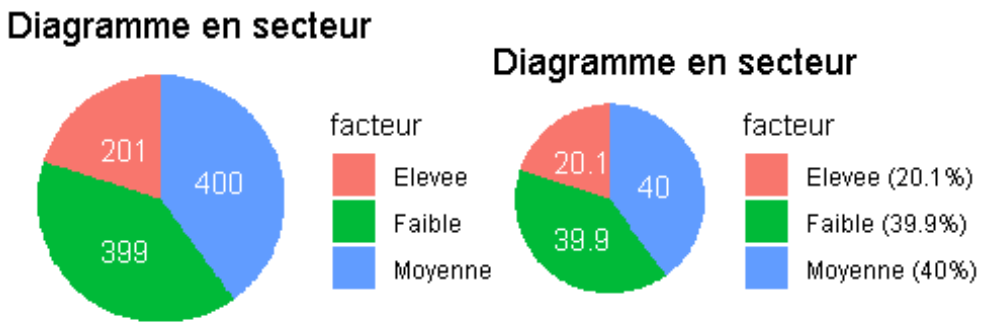
201 livraisons se font dans des conditions élevées et occupent une proportion de 20,1%

399 livraisons se font dans des conditions Faible et occupent une proportion de 39,9%

400 livraisons se font dans des conditions Moyenne et occupent une proportion de 40%

- **Graphique**





Le plus grand nombre de livraisons se fait dans des conditions de circulation Moyenne et occupent 40% des livraisons totales. En revanche, le plus petit nombre de livraisons se fait dans des conditions de circulation Elevée et occupent 20,1% des livraisons totales.

Variable "Time_of_Day"

- **Tableau**

	Effectif	Fréquence
Après-midi	296	0.296
Matin	319	0.319
Nuit	87	0.087
Soir	298	0.298

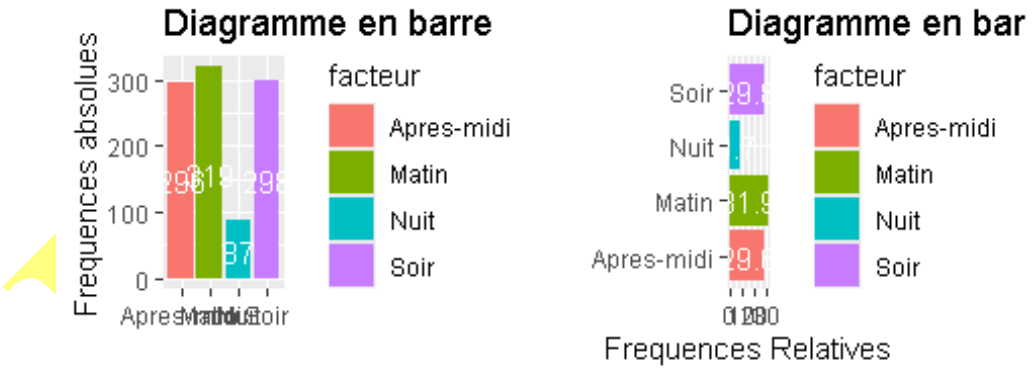
319 livraisons se font le Matin et occupent une proportion de 31,9%

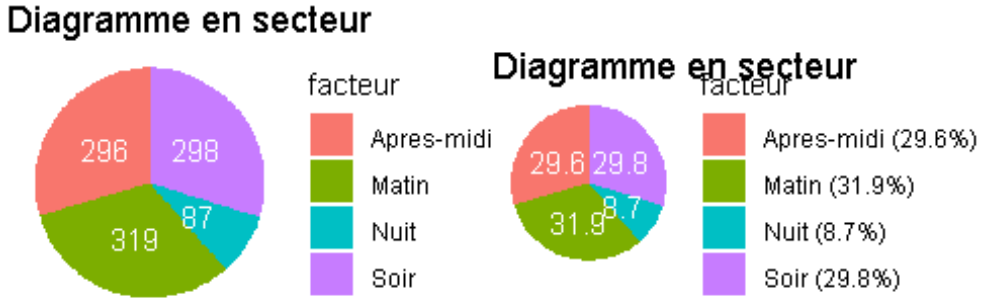
296 livraisons se font l'Après-midi et occupent une proportion de 29,6%

87 livraisons se font la Nuit et occupent une proportion de 8,7%

298 livraisons se font la Nuit et occupent une proportion de 29,8%

- **Graphique**





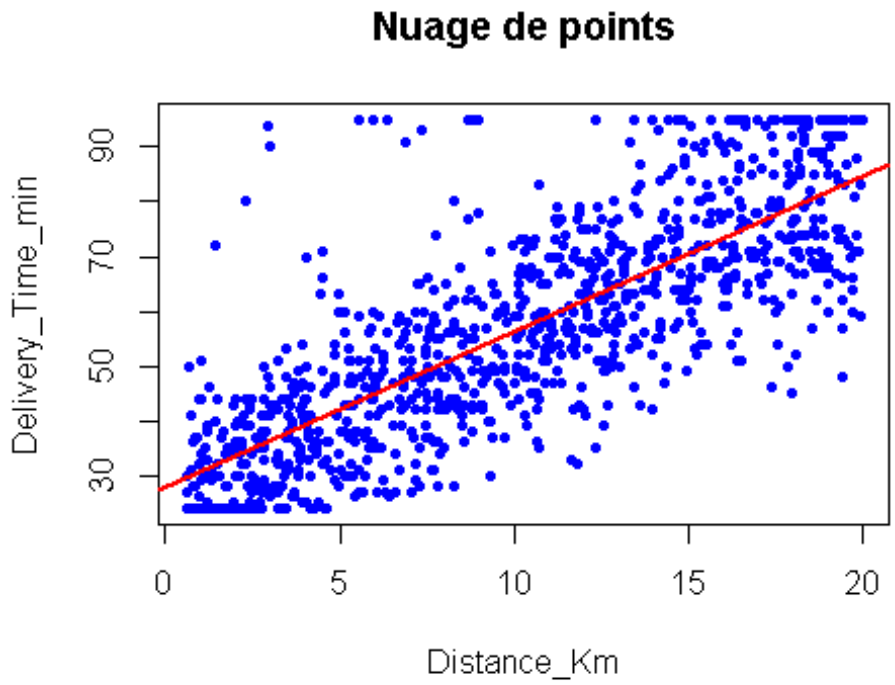
Le plus grand nombre de livraisons se fait le Matin et occupent 31,9% des livraisons totales. En revanche, le plus petit nombre de livraisons se fait la Nuit et occupent 8,7% des livraisons totales.

Cette première partie nous a permis de voir les différentes indicateurs des variables à analyser telles que la variable cible "Delivery_Time_min","Distance_Km","Weather","Traffic_Level","Time_of_Day". Ainsi terminer l'analyse univariée, nous aborderons notre deuxième partie qu'est l'analyse bivariée.

III- ANALYSE BIVARIEE

Variables "Delivery_Time_min" et "Distance_Km"

- **Graphique**



Le nuage de points semble suivre une direction linéaire. Nous avons également remarqué que plus la distance est grande, plus le délai de livraison est long.

- **Equation de la droite**

```
 lm(formula = Delai_livraison$Delivery_Time_min ~ De-
lai_livraison$Distance_km)

 Coefficients:
                 (Intercept)  Delai_livraison$Distance_km
                       27.77                         2.84
```

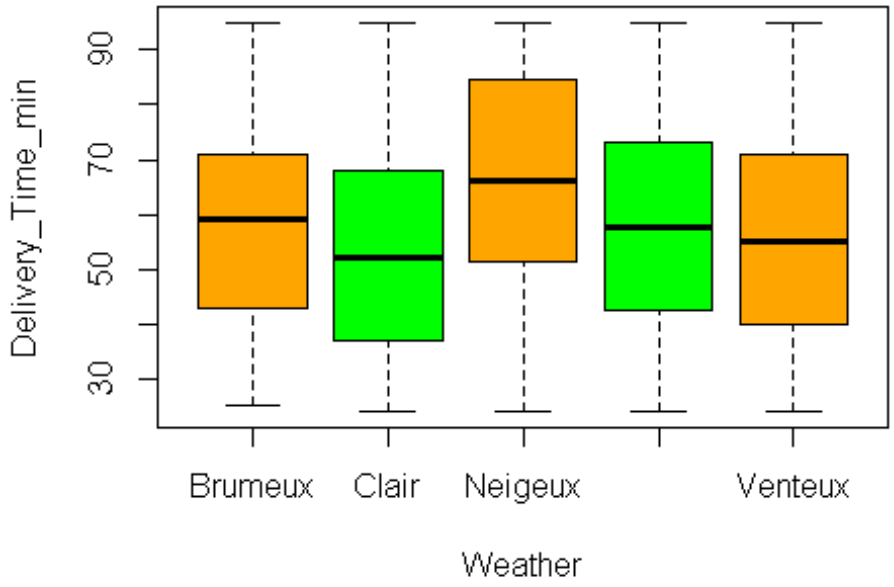
L'équation de la droite est : Delivery_Time_min = 27,77 + 2,84 Distance_Km

- **Liaison**

```
 Correlation_Pearson
 [1] 0.8031169

 Correlation_Spearman
 [1] 0.8167081

 Correlation_Kendall
 [1] 0.6271491

 Coefficient_Determination
 [1] 0.6449968

 Interpretation_Intensite_Liaison
 [1] "liaison forte"

 Coefficents_Droite_Regression
 (Intercept)    vecteur2
   27.768206    2.839551

 Resultat_Test_Liaison

  Pearson's product-moment correlation

 data:  vecteur2 and vecteur1
 t = 42.582, df = 998, p-value < 2.2e-16
 alternative hypothesis: true correlation is not equal to 0
 95 percent confidence interval:
  0.7799561 0.8240809
 sample estimates:
       cor
 0.8031169


 p.value
 [1] 1.15649e-226

 Significacite_Liaison
 [1] "liaison significative"

 Remarque
 [1] "Si la liaison n'est pas significative, Ne pas tenir compte de son in-
tensite"
```

Variables “Delivery_Time_min” et “Météo”

- **Graphique**

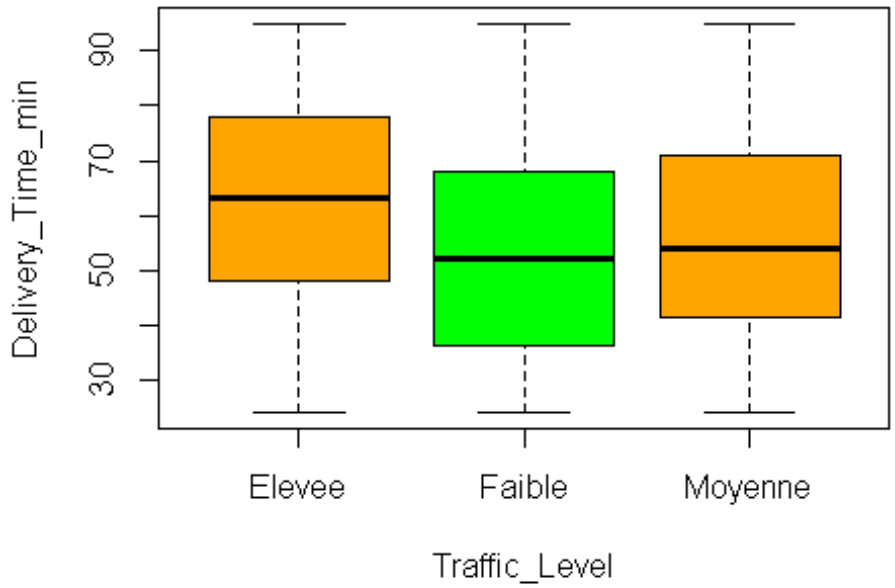


- **Liaison**

```
Rapport_Correlation
 [1] 0.04249462

 Resultat_Test_Anova
 Analysis of Variance Table

 Response: vecteur
            Df Sum Sq Mean Sq F value    Pr(>F)
 facteur     4  17222  4305.5   11.04 9.182e-09 ***
 Residuals 995 388051   390.0
 ---
 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Anova.P.value
 [1] 9.1824e-09

 Significativite_TestAnova
 [1] "liaison significative, les deux variables sont liées"

 Intensite_liaison
 [1] "liaison très faible"

 Remarque
 [1] "Si la liaison n'est pas significative, Ne pas tenir compte de son intensite"
```

Variables “Delivery_Time_min” et “Traffic_Level”

- **Graphique**

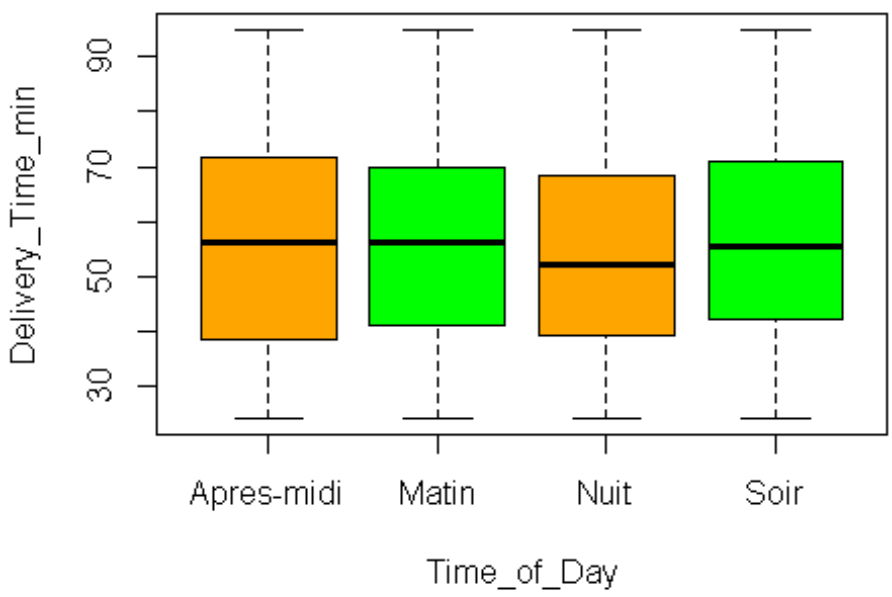


- **Liaison**

```
Rapport_Correlation
[1] 0.03742119

Resultat_Test_Anova
Analysis of Variance Table

Response: vecteur
           Df Sum Sq Mean Sq F value    Pr(>F)
facteur     2  15166  7582.9   19.38 5.533e-09 ***
Residuals 997 390107   391.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Anova.P.value
[1] 5.533433e-09

Significativite_TestAnova
[1] "liaison significative, les deux variables sont liees"

Intensite_liaison
[1] "liaison tr?s faible"

Remarque
[1] "Si la liaison n'est pas significative, Ne pas tenir compte de son intensite"
```

Variables “Delivery_Time_min” et “Time_of_Day”

- **Graphique**



- **Liaison**

```
Rapport_Correlation
 [1] 0.0006992973

 Resultat_Test_Anova
 Analysis of Variance Table

 Response: vecteur
            Df Sum Sq Mean Sq F value Pr(>F)
 facteur     3    283   94.47  0.2323 0.8739
 Residuals 996 404989  406.62

 Anova.P.value
 [1] 0.8738874

 Significativite_TestAnova
 [1] "liaison non significative, les deux variables ne sont pas liees"

 Intensite_liaison
 [1] "liaison tr?s faible"

 Remarque
 [1] "Si la liaison n'est pas significative, Ne pas tenir compte de son intensite"
```

IV- ESTIMATION DE LA MOYENNE DE LA VARIABLE CIBLE “Delivery_Time_min

Tester si la variable suit la loi normale

H0 : la distribution suit une loi normale

H1 : la distribution ne suit pas une loi normale

```
 Shapiro-Wilk normality test

data:  Delai_livraison$Delivery_Time_min
W = 0.9652, p-value = 1.013e-14
```

Conclusion : p-value < 0.05 donc on rejette H0, la distribution ne suit pas une loi normale

Estimation

```
 Bootstrap

data:  Delai_livraison$Delivery_Time_min
1000 replicates

95 percent confidence interval:
 54.980 57.524
sample estimates:
original value
        56.334
```

Conclusion : On a 95% de chance que le vrai Delai moyen de livraison soit compris entre 55 min et 58 min

V- TEST DE CONFORMITE DE LA VARIABLE CIBLE

Conditions

H0 : la distribution suit une loi normale

H1 : la distribution ne suit pas une loi normale

```
 Shapiro-Wilk normality test

data:  Delai_livraison$Delivery_Time_min
W = 0.9652, p-value = 1.013e-14
```

Conclusion : p-value < 0.05 donc on rejette H0, la distribution ne suit pas une loi normale, comme la variable ne suit pas une loi normale nous ferons un test de Wilcoxon

Test

H0 : Delivery_Time_min = 57

H1 : Delivery_Time_min ≠ 57

```
 Wilcoxon signed rank test with continuity correction

data:  Delai_livraison$Delivery_Time_min
V = 228284, p-value = 0.1573
alternative hypothesis: true location is not equal to 57
```

Conclusion : La p-value > 0.05, alors on ne peut rejetter H0. Le delai moyen de livraison n'est pas significativement différent de 57 min

VI- TEST DE LIAISONS

Variable "Delivery_Time" et "Distance_Km"

- **Normalité des variables "Delivery_Time" et "Distance_Km"**

H0 : la distribution suit une loi normale

H1 : la distribution ne suit pas une loi normale

```
 Shapiro-Wilk normality test

data:  Delai_livraison$Delivery_Time_min
W = 0.9652, p-value = 1.013e-14
```

Conclusion : p-value < 0.05 donc on rejette H0, la distribution de la variable "Delivery_Time_min" ne suit pas une loi normale.

H0 : la distribution suit une loi normale

H1 : la distribution ne suit pas une loi normale

```
 Shapiro-Wilk normality test

data:  Delai_livraison$Distance_km
W = 0.95005, p-value < 2.2e-16
```

Conclusion : p-value < 0.05 donc on rejette H0, la distribution de la variable "Distance_Km" ne suit pas une loi normale

Ainsi les variables "Delivery_Time_min" et "Distance_Km" ne suivent pas la loi normale, nous procédérons au test de liaison de Kendall

- **Test de liaison**

H0: rxy = 0 (les variables ne sont pas liées)

H1: rxy ≠ 0 (les variables sont liées)

```
 Kendall's rank correlation tau

data:  Delai_livraison$Delivery_Time_min and Delai_livraison$Distance_km
z = 29.431, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.6271491
```

Conclusion : la p-value < 0.05 donc on rejette H0, le coefficient de corrélation est significativement différent de zéro. C.-à-d. qu'il y a « vraiment » une corrélation entre le Délai de livraison et la distance parcourue. Cette corrélation est estimée à 0.6271491.

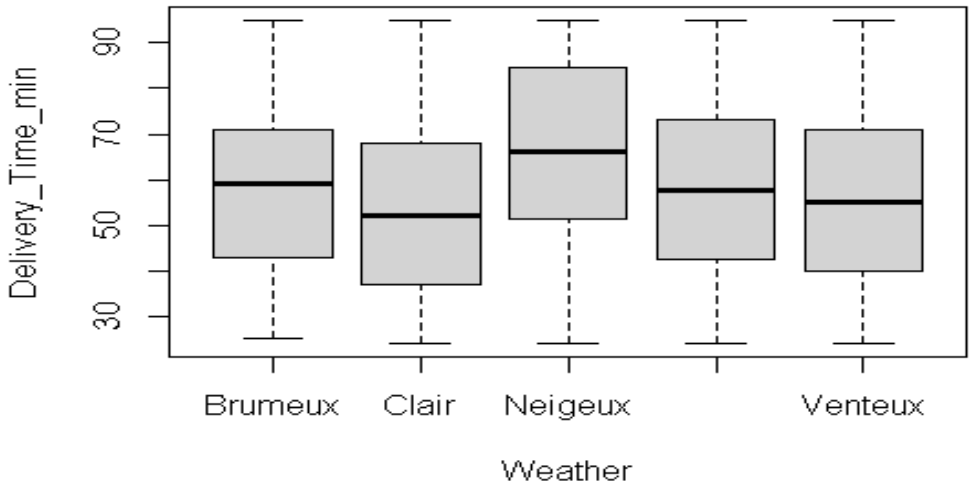
H0:rxy = 0
H1:rxy > 0

```
  Kendall's rank correlation tau

 data:  Delai_livraison$Delivery_Time_min and De-
lai_livraison$Distance_km
 z = 29.431, p-value < 2.2e-16
 alternative hypothesis: true tau is greater than 0
 sample estimates:
       tau
 0.6271491
```

Conclusion : la p-value < 0.05 donc on rejette H0, le coefficient de corrélation est significativement supérieur à zéro. C.-à-d qu'il y a « vraiment » une corrélation positive entre le Délai de livraison et la distance parcourue.

Variable "Delivery_Time" et "Weather"

1ere étape : comparer graphiquement les deux sous population



A travers ce boxplot on peut soupçonner que la variable "Delivery_Time_min" et la variable "Weather" sont liées.

2eme étape : tester la normalité des données dans chaque sous population

H0 : les distributions suivent une loi normale

H1 : les distributions ne suivent pas une loi normale

```
 Shapiro-Wilk normality tests

data:  Delai_livraison$Delivery_Time_min by Delai_livraison$Weather

               W   p-value
Brumeux   0.9639  0.006580 **
Clair     0.9623 7.171e-10 ***
Neigeux   0.9527  0.001335 **
Pluvieux 0.9574 5.847e-06 ***
Venteux   0.9602  0.004934 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On rejette H0 et on conclut que les distributions ne suivent pas une loi normale

3eme étape : tester l'égalité des variances

```
Levene's Test for Homogeneity of Variance (center = median)
        Df F value Pr(>F)
group    4 0.3203 0.8645
      995
```

Conclusion : la p-value > 0.05 donc on ne peut rejeter H0, les variances ne sont pas significativement différentes (on conclut à l'égalité des variances : homoscédasticité)

Etant donné que les variables ne suivent pas la loi normale mais ont des variances égales on fera le test de liaison de Kruskal.wallis

4eme étape : test de liaison

H0 : la variable Delivery_Time_min n'est pas liée à la variable Weather

H1 : la variable Delivery_Time_min est liée à la variable Weather

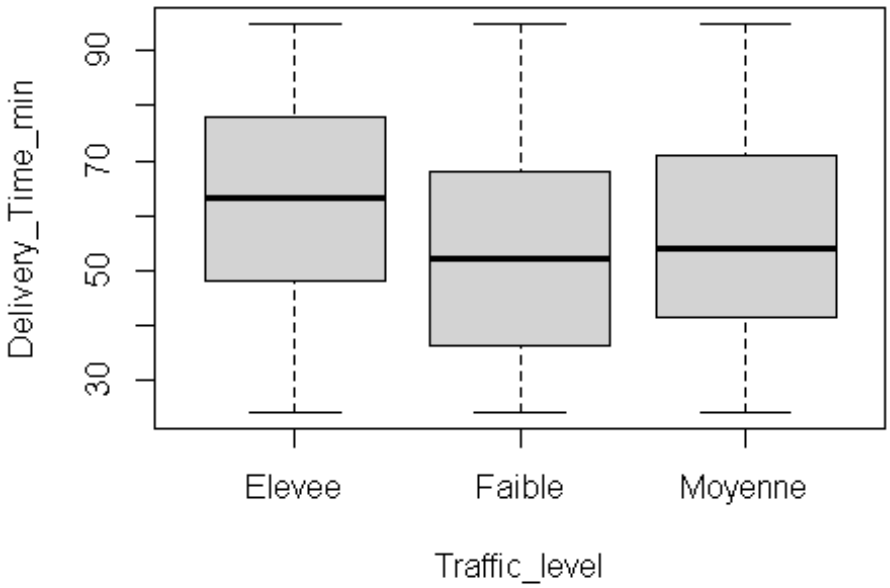
```
 Kruskal-Wallis rank sum test

data:  Delivery_Time_min by Weather
Kruskal-Wallis chi-squared = 38.477, df = 4, p-value = 8.931e-08
```

Conclusion : la p-value < 0.05 donc on rejette H0, Le rapport de corrélation est significativement différent de zéro. C.-à-d. que les deux

variables ne sont pas indépendantes : il y une liaison entre Delivery_Time_min et Weather.

Variable “Delivery_Time_min” et “Traffic_level”

1ere étape : comparer graphiquement les deux sous population



A travers ce boxplot on peut soupçonner que la variable “Delivery_Time_min” et la variable “Traffic_level” sont liées.

2eme étape : tester la normalité des données dans chaque sous population

H0 : les distributions suivent une loi normale

H1 : les distributions ne suivent pas une loi normale

```
  Shapiro-Wilk normality tests

 data:  Delai_livraison$Delivery_Time_min by De-
lai_livraison$Traffic_Level
```

```
             W   p-value
Elevee   0.9606 2.174e-05 ***
Faible   0.9599 5.729e-09 ***
Moyenne 0.9658 4.825e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On rejette H0 et on conclut que les distributions ne suivent pas une loi normale

3eme étape : tester l'égalité des variances

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   2  0.0012 0.9988
      997
```

Conclusion : la p-value > 0.05 donc on ne peut rejeter H0, les variances ne sont pas significativement différentes (on conclut à l'égalité des variances : homoscédasticité)

4eme étape : test de liaison

H0 : la variable Delivery_Time_min n'est pas liée à la variable Traffic_level

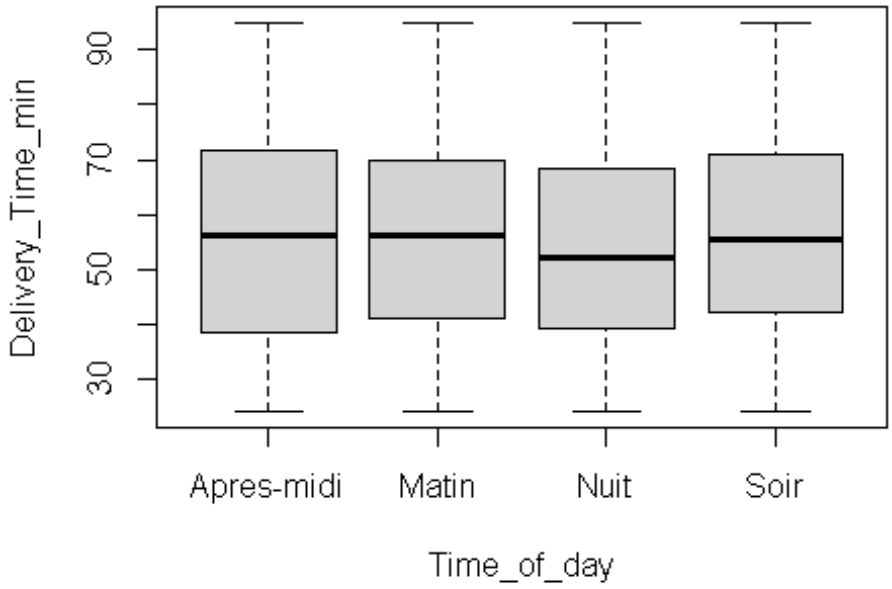
H1 : la variable Delivery_Time_min est liée à la variable Traffic_level

```
 Kruskal-Wallis rank sum test

data:  Delivery_Time_min by Traffic_Level
Kruskal-Wallis chi-squared = 35.144, df = 2, p-value = 2.337e-08
```

Conclusion : la p-value < 0.05 donc on rejette H0, Le rapport de corrélation est significativement différent de zéro. C.-à-d. que les deux variables ne sont pas indépendantes : il y une liaison entre Delivery_Time_min et Traffic_level.

Variable “Delivery_Time_min” et “Time_of_day”

1ere étape : comparer graphiquement les deux sous population



A travers ce boxplot on peut soupçonner que la variable “Delivery_Time_min” et la variable “Time_of_day” ne sont pas liées.

2eme étape : Tester la normalité des données dans chaque sous population

H0 : les distributions suivent une loi normale

H1 : les distributions ne suivent pas une loi normale

```
  Shapiro-Wilk normality tests

 data:  Delai_livraison$Delivery_Time_min by De-
lai_livraison$Time_of_Day

                 W   p-value
 Apres-midi 0.9594 2.397e-07 ***
 Matin      0.9648 5.459e-07 ***
 Nuit       0.9486  0.001698 **
```

```
Soir        0.9696 6.129e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On rejette H0 et on conclut que les distributions ne suivent pas une loi normale

3eme étape : Tester l'égalité des variances

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   3  0.5225 0.6669
      996
```

Conclusion : la p-value > 0.05 donc on ne peut rejeter H0, les variances ne sont pas significativement différentes (on conclut à l'égalité des variances : homoscédasticité)

4eme étape : Test de liaison

H0 : la variable Delivery_Time_min n'est pas liée à la variable Traffic_level

H1: la variable Delivery_Time_min est liée à la variable Time_of_day

```
 Kruskal-Wallis rank sum test

data:  Delivery_Time_min by Time_of_Day
Kruskal-Wallis chi-squared = 0.96505, df = 3, p-value = 0.8097
```

Conclusion : la p-value > 0.05 donc on ne peut rejeter H0, le rapport de corrélation n'est pas significativement différent de zéro. C.-à-d. que les deux variables sont indépendantes : il n'y pas de liaison entre Delivery_Time_min et Time_of_day.

VII- MODELISATION REGRESSION MULTIPLE

1ere étape : Importer les données

```
 'data.frame':    1000 obs. of  5 variables:
  Delivery_Time_min: num  43 84 59 37 68 57 49 46 35 73 ...
  Distance_km      : num  7.93 16.42 9.52 7.44 19.03 ...
  Weather          : Factor w/ 5 levels "Brumeux","Clair",..: 5 2 1 4
2 2 2 2 3 1 ...
  Traffic_Level    : Factor w/ 3 levels "Elevee","Faible",..: 2 3 2 3
2 2 2 3 2 2 ...
  Time_of_Day      : Factor w/ 4 levels "Apres-midi","Matin",..: 1 4
3 1 2 4 2 4 4 4 ...
```

Specification du modele: Delivery_Time_min = B0 + B1Distance_km + B2Weather + B3Traffic_Level + B4Time_of_Day +ε

2eme étape : Estimer les paramètres

- **Faire la regression**

```
             (Intercept)          df$Distance_km        df$WeatherClair
              38.6901989              2.8028294             -6.6300214
      df$WeatherNeigeux      df$WeatherPluvieux       df$WeatherVenteux
               2.1798003             -2.4313371             -5.7552902
 df$Traffic_LevelFaible df$Traffic_LevelMoyenne     df$Time_of_DayMatin
             -10.1773162             -5.4376325             -1.0935763
     df$Time_of_DayNuit      df$Time_of_DaySoir
              -0.6973264              0.6467414
```

- **Voir l'intervalle de confiance des coefficients estimés**

```
                              2.5 %     97.5 %
(Intercept)              35.6826772 41.6977206
df$Distance_km            2.6818279  2.9238308
df$WeatherClair          -8.9848025 -4.2752402
df$WeatherNeigeux        -0.8776233  5.2372238
df$WeatherPluvieux       -5.0376984  0.1750242
df$WeatherVenteux        -8.8373071 -2.6732733
df$Traffic_LevelFaible  -12.0554953 -8.2991371
df$Traffic_LevelMoyenne  -7.3162030 -3.5590619
df$Time_of_DayMatin      -2.8447408  0.6575882
df$Time_of_DayNuit       -3.3504272  1.9557744
df$Time_of_DaySoir       -1.1379431  2.4314259
```

3eme étape : Test de significativité Globale

H0 : B0 = B1 = … = Bn

H1: B0 ≠ B1 ≠ … ≠ Bn

```
Call:
lm(formula = df$Delivery_Time_min ~ df$Distance_km + df$Weather +
    df$Traffic_Level + df$Time_of_Day)

Residuals:
    Min      1Q  Median      3Q     Max
-30.010  -7.056  -0.502   6.269  60.259

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               38.69020    1.53260  25.245  < 2e-16 ***
df$Distance_km             2.80283    0.06166  45.455  < 2e-16 ***
df$WeatherClair           -6.63002    1.19997  -5.525 4.21e-08 ***
df$WeatherNeigeux          2.17980    1.55803   1.399 0.162104
df$WeatherPluvieux        -2.43134    1.32817  -1.831 0.067463 .
df$WeatherVenteux         -5.75529    1.57056  -3.664 0.000261 ***
df$Traffic_LevelFaible   -10.17732    0.95710 -10.633  < 2e-16 ***
df$Traffic_LevelMoyenne   -5.43763    0.95730  -5.680 1.77e-08 ***
df$Time_of_DayMatin       -1.09358    0.89237  -1.225 0.220691
df$Time_of_DayNuit        -0.69733    1.35199  -0.516 0.606125
df$Time_of_DaySoir         0.64674    0.90946   0.711 0.477171
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.04 on 989 degrees of freedom
Multiple R-squared:  0.7028, Adjusted R-squared:  0.6998
F-statistic: 233.9 on 10 and 989 DF,  p-value: < 2.2e-16
```

P-value < 0,05 donc on rejette H0 et on conclut que le modèle est globalement significatif. Puisque le modele est significatif, on peut s'intéresser directement au coefficient de détermination. La valeur de R2 est donnée, ainsi que le R2 ajustée. La valeur du R2 est très élevée (R2 = 0.7033). En d'autres termes,70,33% de la variabilité du Délai de livraison est expliquée par l'ensemble des explicatives.

4eme étape : Test de significativité individuel

```
Analysis of Variance Table

Response: df$Delivery_Time_min
                 Df Sum Sq Mean Sq  F value    Pr(>F)
df$Distance_km     1 261399  261399 2146.557 < 2.2e-16 ***
df$Weather         4   8844    2211   18.157 2.104e-14 ***
```

```
df$Traffic_Level    2  14093    7046   57.864 < 2.2e-16 ***
df$Time_of_Day      3    499     166    1.366    0.2517
Residuals         989 120437     122
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ainsi les variables Distance_Km, Weather, Traffic_level sont significatif et peuvent être sujet à interprétation. Déterminions les coefficients estimés des variables et interprétons ceux qui sont significatifs.

```
Call:
lm(formula = df$Delivery_Time_min ~ df$Distance_km + df$Weather +
    df$Traffic_Level + df$Time_of_Day)

Residuals:
    Min      1Q  Median      3Q     Max
-30.010  -7.056  -0.502   6.269  60.259

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               38.69020    1.53260  25.245  < 2e-16 ***
df$Distance_km             2.80283    0.06166  45.455  < 2e-16 ***
df$WeatherClair           -6.63002    1.19997  -5.525 4.21e-08 ***
df$WeatherNeigeux          2.17980    1.55803   1.399 0.162104
df$WeatherPluvieux        -2.43134    1.32817  -1.831 0.067463 .
df$WeatherVenteux         -5.75529    1.57056  -3.664 0.000261 ***
df$Traffic_LevelFaible   -10.17732    0.95710 -10.633  < 2e-16 ***
df$Traffic_LevelMoyenne   -5.43763    0.95730  -5.680 1.77e-08 ***
df$Time_of_DayMatin       -1.09358    0.89237  -1.225 0.220691
df$Time_of_DayNuit        -0.69733    1.35199  -0.516 0.606125
df$Time_of_DaySoir         0.64674    0.90946   0.711 0.477171
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.04 on 989 degrees of freedom
Multiple R-squared:  0.7028, Adjusted R-squared:  0.6998
F-statistic: 233.9 on 10 and 989 DF,  p-value: < 2.2e-16
```

INTERPRETATION

- **Ici l'intercept (ou la constante) est estimé à 57.61464 avec un P-value < 0.05, ce qui veut dire la constante est significativement différent de zéro et qu'il doit apparaitre dans le modèle. La va-**

leur moyenne du Délai du livraison quand les explicatives sont nulles est égale à 39 minutes.

- **Ici le coefficient estimé de la Distance est 0.14430 avec un P-value < 0.05, ce qui veut dire ce coefficient est significativement différent de zéro. Cela indique qu'il y a une liaison significative entre le Délai de livraison et la Distance parcourue. Le Délai de livraison moyen augmente de 2,8 minutes pour chaque Distance supplémentaire parcourue.**

- **(Brumeux pris ici comme modalité de référence), Par rapport à un Délai de livraison dans un temps Brumeux, un Délai de livraison dans un temps Clair à 6,6 minutes de moins.**

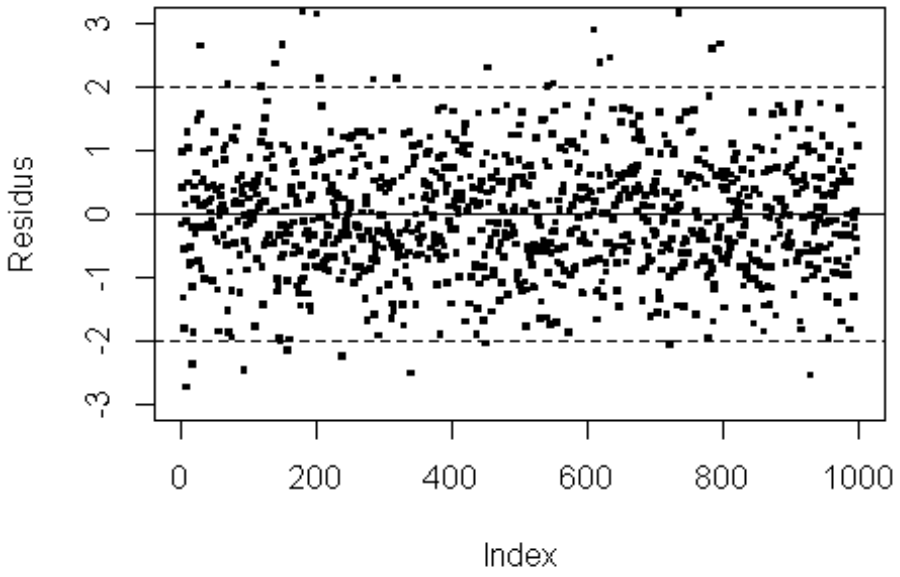
- **(Brumeux pris ici comme modalité de référence), Par rapport à un Délai de livraison dans un temps Brumeux, un Délai de livraison dans un temps Venteux à 5,9 minutes de moins.**

- **(Elevee pris ici comme modalité de référence), Par rapport à un Délai de livraison dans une condition de circulation Elevee, un Délai de livraison dans une condition de circulation Faible prend 10,3 minutes de moins.**

- **(Elevee pris ici comme modalité de référence), Par rapport à un Délai de livraison dans une condition de circulation Elevee, un Délai de livraison dans une condition de circulation Moyenne prend 5,4 minutes de moins.**

5eme étape : Analyse des résidus

- **Graphique**



En théorie 95% des résidus studentisées se trouvent dans l'intervalle [-2;2]. Ici on a visuellement beaucoup de résidus qui se trouvent dans cet intervalle, ce qui est acceptable.

```
[1] 96
```

6eme étape : Analyse de la Multicolinéarité

Mais avant voyons quel modèle allons-nous choisir pour effectuer la multi colinéarité.

a) Modele avec la variable Time_of_Day

```
R² ajusté : 0.6998207
AIC : 7653.001
BIC : 7711.894
Test F (valeur, df1, df2) : 233.9011 10 989
p-value du test F : 1.578568e-252
```

b) Modele sans la variable Time of day

```
R² ajusté : 0.6994885
AIC : 7651.136
BIC : 7695.305
Test F (valeur, df1, df2) : 333.1903 7 992
p-value du test F : 2.195346e-255
```

- **Le R^2 ajusté diminue légèrement après le retrait de Time of Day (de 0.7003243 à 0.6998286). Cette diminution est très faible, ce qui suggère que Time of Day n'apporte qu'une contribution marginale à l'explication de la variance de la variable dépendante.**

- **L'AIC diminue légèrement après le retrait de Time of Day (de 7651.322 à 7650.003). Une diminution de l'AIC indique que le modèle sans Time of Day est légèrement meilleur en termes de qualité d'ajustement et de parcimonie.**

- **Le BIC diminue de manière plus significative après le retrait de Time of Day (de 7710.215 à 7694.173). Le BIC pénalise plus fortement les modèles complexes, donc cette diminution suggère que le modèle sans Time of Day est préférable.**

- **Test F global Les deux modèles sont globalement significatifs (p-value du test F très faible dans les deux cas). Cela signifie que les variables explicatives, dans leur ensemble, ont un effet significatif sur la variable dépendante, que Time_of_day soit incluse ou non.**

D'après ces conclusions, nous décidons de choisir le modèle avec la variable "Time_of_day" à des fins prédictifs et le modèle sans la variable "Time_of_day" à des fins explicatives.

c) Multicolinéarité

	Variables	Tolerance	VIF
1	df$Distance_km	0.9894972	1.010614
2	df$WeatherClair	0.3391945	2.948161
3	df$WeatherNeigeux	0.5629470	1.776366
4	df$WeatherPluvieux	0.4141782	2.414420
5	df$WeatherVenteux	0.5662492	1.766007
6	df$Traffic_LevelFaible	0.5552488	1.800994
7	df$Traffic_LevelMoyenne	0.5545651	1.803215

La plupart des variables n'ont pas de problème apparent de Multicolinéarité, ce qui est bien. Les valeurs de tolérance et de VIF suggèrent que les variables expliquent des aspects différents du délai de livraison, sans redondance significative. Cependant, les variables météorologiques telles que Weather$Clair et Weather$Pluvieux montrent une légère indication de multicolinéarité. Cela peut nécessiter une surveillance, mais ne devrait pas être une source majeure de préoccupation.

8eme étape : Validation du modèle

1) Test de linéarité du modèle

H0 : Le modèle est linéaire

H1 : Le modèle n'est pas linéaire

```
data:  regB
Rain = 1.1514, df1 = 500, df2 = 492, p-value = 0.05844
```

P-value < 0,05 donc on conclut que modele n'est pas linéaire

- **Regression avec interaction**

```
 Rainbow test

data:  regC
Rain = 1.1371, df1 = 500, df2 = 484, p-value = 0.07744
```

P-value > 0,05 donc on conclut que ce modele est linéaire

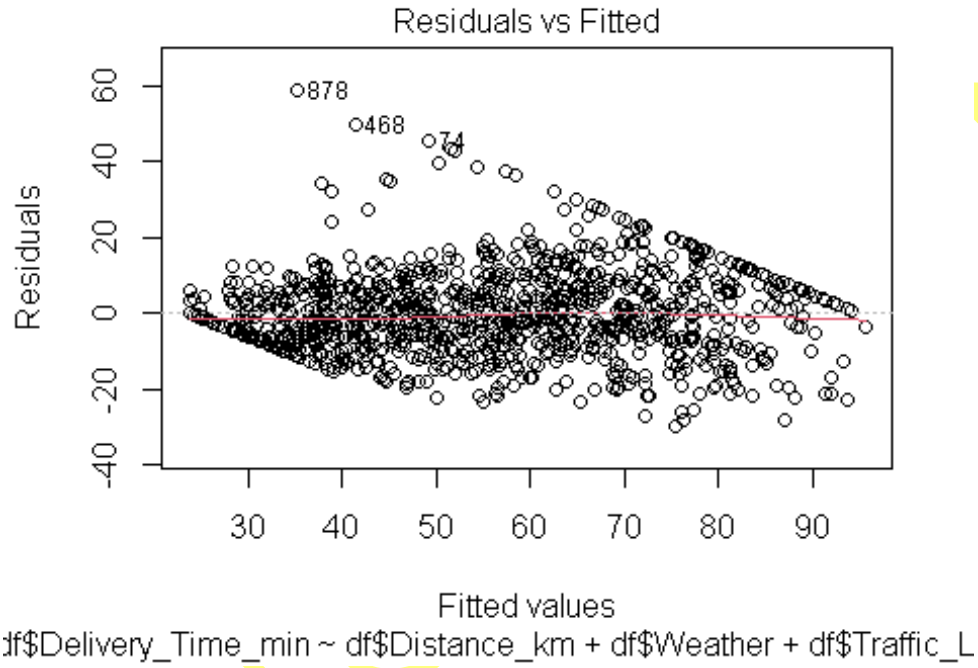
Le problème ici avec le modele avec interaction est qu'il y a Multicolinéarité car la plupart des vifs sont supérieur à 5 et les tolérances sont supérieur à 0,02 :

	Variables	Tolerance	VIF
1	df$Distance_km	0.97270398	1.028062
2	df$WeatherClair	0.06824631	14.652808
3	df$WeatherNeigeux	0.11459546	8.726349
4	df$WeatherPluvieux	0.08564557	11.676028
5	df$WeatherVenteux	0.12512441	7.992046
6	df$Traffic_LevelFaible	0.06089914	16.420594
7	df$Traffic_LevelMoyenne	0.05400158	18.517978
8	df$WeatherClair:df$Traffic_LevelFaible	0.07728314	12.939433

```
 9    df$WeatherNeigeux:df$Traffic_LevelFaible 0.15972264  6.260853
10  df$WeatherPluvieux:df$Traffic_LevelFaible 0.12228226  8.177801
11   df$WeatherVenteux:df$Traffic_LevelFaible 0.24881578  4.019038
12    df$WeatherClair:df$Traffic_LevelMoyenne 0.06487134 15.415127
13  df$WeatherNeigeux:df$Traffic_LevelMoyenne 0.20395277  4.903096
14 df$WeatherPluvieux:df$Traffic_LevelMoyenne 0.12349410  8.097553
15  df$WeatherVenteux:df$Traffic_LevelMoyenne 0.15977866  6.258658
```

2) Homoscédasticité des érreurs

- **Graphique**



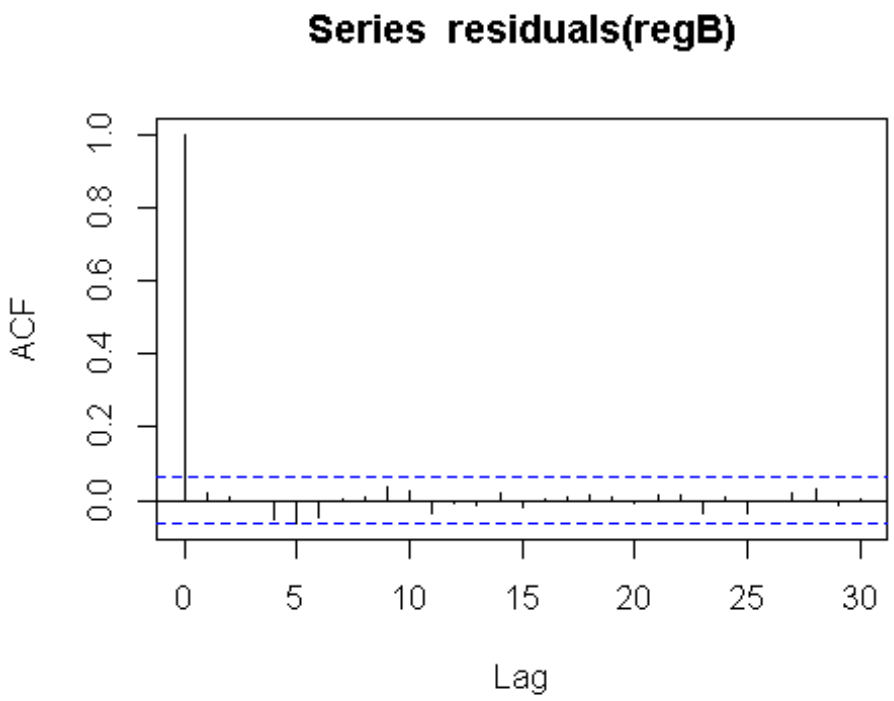
- **TEST BREUSCH-PAGAN**
  H0 : il y'a homoscédasticité
  H1 : il y'a hétéroscédasticité

```
	studentized Breusch-Pagan test

data:  regB
BP = 30.913, df = 15, p-value = 0.009024
```

P-value = 0.007536 < 0,05 donc on rejette H0 et on conclut qu'il y'a hétéroscédasticité

3) Autocorrelation des érreurs

- **Graphique**



- **TEST DE DURBIN-WATSON**

H0 : il n'y a pas autocorrélation

H1: il y a autocorrelation

```
 Durbin-Watson test

data:  regB
DW = 1.9596, p-value = 0.2609
alternative hypothesis: true autocorrelation is greater than 0
```

P-value = 0.2482 > 0,05 donc on ne peut rejeter H0 et on conclut qu'il n'y a pas autocorrélation d'ordre 1

4) Normalité des érreurs

H0 : La distribution suit une loi normale

H1 : La distribution ne suit pas une loi normale

```
 Shapiro-Wilk normality test

data:  residuals(regB)
W = 0.97214, p-value = 6.337e-13
```

P-value = 5.254e-13 < 0,05 donc on rejette H0 et on conclut que la distribution des erreurs ne suit pas une loi normale.

Ainsi, d'après 1),2),3),4) on peut s'apercevoir que les hypothèses des MCO sont violées et on conclut que le modèle n'est pas valide. Cependant, il peut exister des méthodes de régression plus performantes telles que les MCG, la régression polynomiale, les modèles de machine Learning etc….

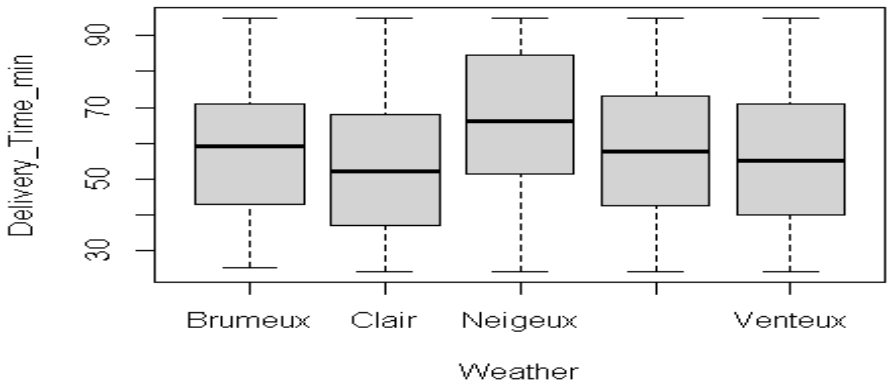
VIII- MODELISATION ANOVA-1

On veut modéliser Le "Délai de livraison" en fonction de la Météo par un modèle d'ANOVA :

Delivery_Time_min = Delivery_moyen + Delivery_Brumeux + Delivery_Clair + Delivery_Neigeux + Delivery_Pluvieux + Delivery_Venteux + Eij

1ere étape : comparer graphiquement les sous populations

- **Graphique**



De manière générale, lorsque les conditions météorologiques se dégradent, on peut affirmer que les délais de livraison s'allongent par rapport à ceux observés en l'absence de perturbations météorologiques. On peut donc raisonnablement supposer que les délais de livraison sont intrinsèquement liés à l'état de la météo.

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

- **Moyenne(mean)**

```
  Brumeux    Clair  Neigeux Pluvieux  Venteux
 58.87379 52.97546 66.19192 58.79245 55.13402
```

- **Quantile**

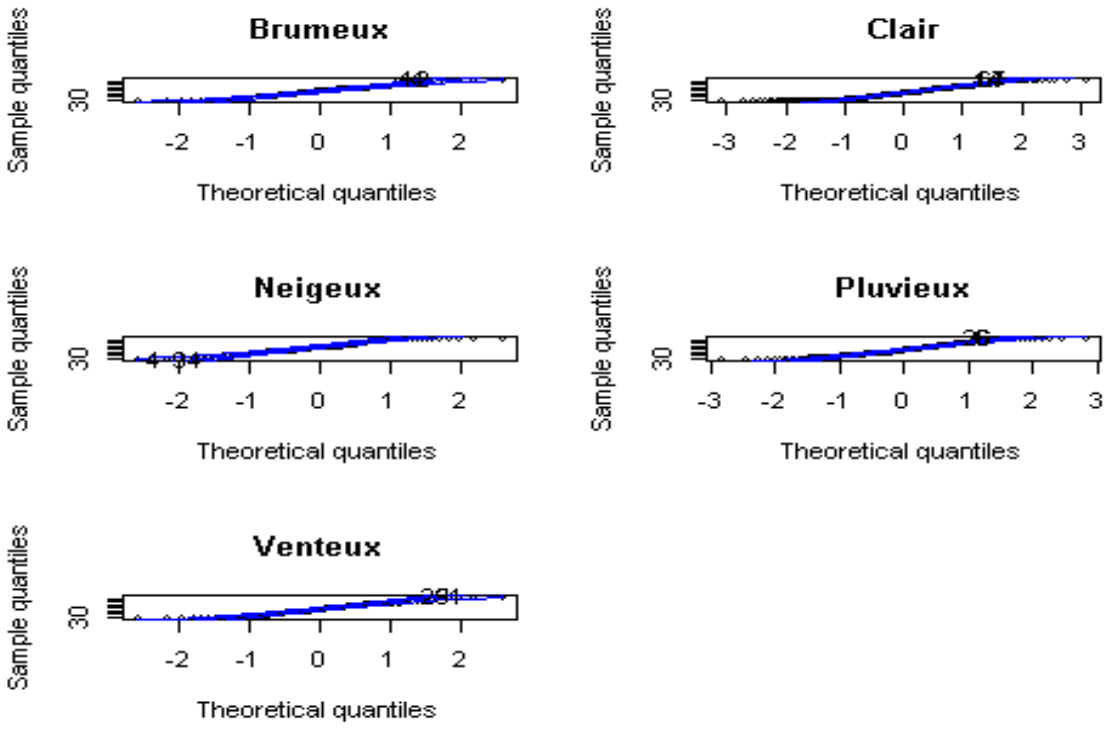
```
Brumeux
   0%  25%  50%  75% 100%
   25   43   59   71   95

 Clair
   0%  25%  50%  75% 100%
   24   37   52   68   95

 Neigeux
   0%  25%  50%  75% 100%
 24.0 51.5 66.0 84.5 95.0

 Pluvieux
    0%   25%   50%   75%  100%
 24.00 42.75 57.50 73.00 95.00

 Venteux
   0%  25%  50%  75% 100%
   24   40   55   71   95
```

- **Ecart-type(sd)**

```
  Brumeux    Clair  Neigeux Pluvieux  Venteux
 19.59025 19.53736 19.51696 20.24458 20.11044
```

3eme étape : Tester la normalité des données dans chaque sous population

- **Graphique**



- **Test de normalité**

H0 : la distribution suit une loi normale

H1 : la distribution ne suit pas une loi normale

```
 Shapiro-Wilk normality tests

data:  Delai_livraison$Delivery_Time_min by Delai_livraison$Weather

               W   p-value
Brumeux   0.9639  0.006580 **
Clair     0.9623 7.171e-10 ***
Neigeux   0.9527  0.001335 **
Pluvieux  0.9574 5.847e-06 ***
Venteux   0.9602  0.004934 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion : les p-value < 0.05 donc on rejette H0, on conclut que les distributions ne suivent pas la loi normale

4eme étape : Tester l'égalité des variances

```
 Kruskal-Wallis rank sum test

data:  Delai_livraison$Delivery_Time_min by Delai_livraison$Weather
Kruskal-Wallis chi-squared = 38.477, df = 4, p-value = 8.931e-08
```

Conclusion : la p-value < 0.05 donc on rejette H0, Au moins une des moyennes est significativement différente des autres.

5eme étape : faire un test robuste par Bootstrap (rééchantil-lonnage)

```
Monte-Carlo test

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

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

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 Délai de livraison.

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

H0 : μ1=μ2=…=μn Toutes les moyennes sont égales

H1 : μi ≠ μj Au moins une des moyennes est différente des autres

```
Analysis of Variance Table

Response: Delai_livraison$Delivery_Time_min
                          Df Sum Sq Mean Sq F value    Pr(>F)
Delai_livraison$Weather    4  17222  4305.5   11.04 9.182e-09 ***
Residuals                995 388051   390.0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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 Délai de livraison.

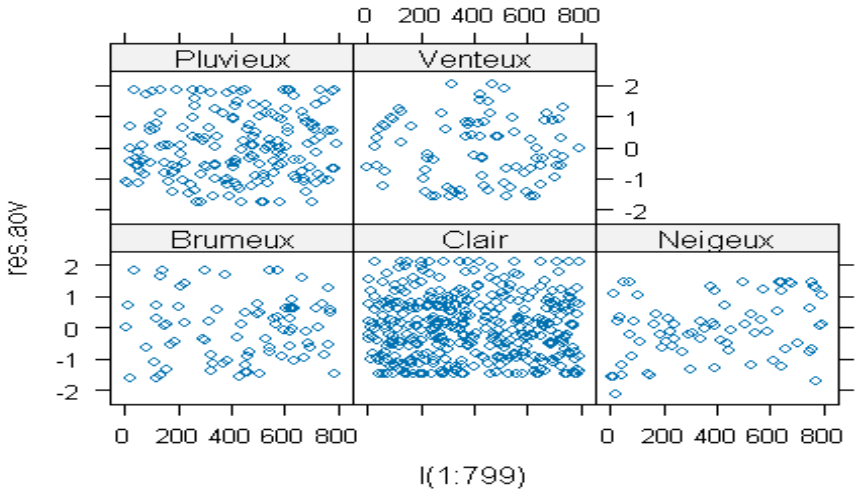
Il est dans ce cas nécessaire de réaliser des comparaisons deux-à-deux pour, identifier les classes en question.

```
 Pairwise comparisons using permutation t tests

data:  Delai_livraison$Delivery_Time_min and Delai_livraison$Weather
999 permutations

          Brumeux Clair  Neigeux Pluvieux
Clair     0.0160  -      -       -
Neigeux   0.0200  0.0067 -       -
Pluvieux  0.9940  0.0067 0.0100  -
Venteux   0.1825  0.3733 0.0067  0.1825

P value adjustment method: fdr
```

Conclusion : on peut ici soupçonner les moyennes (Pluvieux et Brumeux) d'être responsables du rejet de l'hypothèse nulle.

7eme étape : Analyser les résidus

- **Graphique**



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

```
 [1] 97.9
```

8eme étape : Interpréter les coefficients

Lors du test de la significativité du facteur, nous avons constaté qu'il y a un effet global de la météo sur le Délai de livraison :

```
 Call:
 lm(formula = Delai_livraison$Delivery_Time_min ~ De-
lai_livraison$Weather)

 Residuals:
     Min       1Q   Median       3Q      Max
 -42.192 -15.975  -0.975  14.823  42.025

 Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
 (Intercept)                    58.87379    1.94587  30.256  < 2e-16
***
 Delai_livraison$WeatherClair   -5.89833    2.14102  -2.755  0.00598
**
 Delai_livraison$WeatherNeigeux  7.31813    2.77953   2.633  0.00860
**
 Delai_livraison$WeatherPluvieux -0.08133    2.37193  -0.034  0.97265
 Delai_livraison$WeatherVenteux  -3.73977    2.79411  -1.338  0.18106
 ---
 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Residual standard error: 19.75 on 995 degrees of freedom
 Multiple R-squared:  0.04249,    Adjusted R-squared:  0.03865
 F-statistic: 11.04 on 4 and 995 DF,  p-value: 9.182e-09
```

- **En moyenne le délai de livraison dans un temps Brumeux est 58 minutes.**
- **Par rapport à une livraison qui se fait dans un temps Brumeux, une livraison qui se fait dans un temps Clair prend 6 minutes de moins.**

- **Par rapport à une livraison qui se fait dans un temps Brumeux, une livraison qui se fait dans un temps Neigeux prend 7 minutes de plus.**

Si on prend la contrainte ($\sum \alpha i = 0$) ce qui revient à prendre la moyenne comme référence.

```
Call:
lm(formula = Delivery_Time_min ~ C(Weather, sum), data = De-
lai_livraison)

Residuals:
    Min      1Q  Median      3Q     Max
-42.192 -15.975  -0.975  14.823  42.025

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        58.3935     0.7585  76.984  < 2e-16 ***
C(Weather, sum)1    0.4803     1.6874   0.285    0.776
C(Weather, sum)2   -5.4181     1.0266  -5.278 1.61e-07 ***
C(Weather, sum)3    7.7984     1.7143   4.549 6.06e-06 ***
C(Weather, sum)4    0.3989     1.2958   0.308    0.758
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.75 on 995 degrees of freedom
Multiple R-squared:  0.04249,    Adjusted R-squared:  0.03865
F-statistic: 11.04 on 4 and 995 DF,  p-value: 9.182e-09
```

μ = 58.2931 est le délai de livraison moyen de tout l'échantillon, le dernier coefficient se déduit de l'équation : $\sum \alpha i = 0$

$0.5807 - 5.1783 + 7.5869 + 0.3752 + \alpha 5 = 0$ donc $\alpha 5 = -3.3645$

Si on fait le test de significativité :

H0 : $\alpha 5 = 0$

H1 : $\alpha 5 \neq 0$

```
alpha5: -3.3645

SE(alpha5): 4.660711

t-value: -0.7218856

p-value: 0.4705344
```

P-value :0.4708337 > 0,05 alors On ne peut rejeter H0 et on conclut que l'effet propre de la modalité 5 n'est pas significatif.

INTERPRETATION

- **α3 = -5.1783, cela signifie que, par rapport à la moyenne globale du temps de livraison, les livraisons effectuées par temps clair sont en moyenne 5.1783 minutes plus courtes.**
- **α4 = 7.5869, Cela signifie que, par rapport à la moyenne globale du temps de livraison, les livraisons effectuées par temps neigeux sont en moyenne 7.5869 minutes plus longues.**

Les conditions météorologiques jouent un rôle crucial dans l'efficacité des livraisons. Alors que le temps clair facilite des livraisons plus rapides, le temps neigeux entraîne des retards significatifs. Ces résultats soulignent l'importance de prendre en compte les conditions météorologiques dans la planification logistique et la gestion des attentes des clients. Une meilleure préparation aux conditions météorologiques extrêmes pourrait grandement améliorer la performance globale des services de livraison.

CONCLUSION GENERALE DE L'ETUDE

Cette étude visait à modéliser le délai de livraison en fonction des conditions météorologiques l'aide d'une régression linéaire multiple et d'une analyse ANOVA. Les résultats ont montré que les conditions météorologiques ont un impact significatif sur le temps de livraison. Plus précisément, les livraisons effectuées par temps clair sont en moyenne 5.18 minutes plus rapides que la moyenne globale, tandis que celles effectuées par temps neigeux sont 7.59 minutes plus longues. Ces effets sont statistiquement significatifs, ce qui confirme l'importance des conditions météorologiques dans la performance des services de livraison. Cependant, l'étude a également révélé que les hypothèses des Moindres Carrés Ordinaires (MCO) pourraient être violées, ce qui remet en question la validité du modèle de régression linéaire multiple. Des violations telles que la non-linéarité, l'hétéroscédasticité ou l'autocorrélation des erreurs pourraient affecter la fiabilité des résultats. Par conséquent, bien que les conclusions sur l'impact des conditions météorologiques soient pertinentes, il est essentiel d'explorer des modèles alternatifs et d'améliorer la méthodologie pour garantir des résultats robustes.

RECOMMANDATIONS

1.Vérification des hypothèses des MCO :

- **Effectuer des diagnostics approfondis pour vérifier les hypothèses des MCO (linéarité, homoscédasticité, indépendance et normalité des erreurs).**
- **Utiliser des outils tels que les graphiques des résidus, le test de Breusch-Pagan (hétéroscédasticité) et le test de Durbin-Watson (autocorrélation).**

2.Exploration de modèles alternatifs:

- **En cas de violation des hypothèses, envisager des méthodes alternatives telles que : Moindres Carrés Généralisés (MCG) pour gérer l'hétéroscédasticité ou l'autocorrélation.**
- **Régression polynomiale pour capturer des relations non linéaires.**

- **Modèles de machine Learning (forêts aléatoires, réseaux de neurones, régression LASSO/Ridge) pour mieux modéliser des données complexes.**

3.Inclusion de variables supplémentaires :

Intégrer d'autres facteurs explicatifs potentiels, tels que : - La distance de livraison. - Le trafic routier. - L'heure de la journée. - Le type de véhicule utilisé. - Cela permettrait de capturer davantage de variabilité dans les délais de livraison et d'améliorer la précision du modèle.

4.Amélioration de la planification logistique :

- **Utiliser les résultats pour optimiser les plannings de livraison en fonction des prévisions météorologiques.**
- **Allouer plus de ressources (véhicules, personnel) pendant les périodes de mauvais temps pour minimiser les retards.**

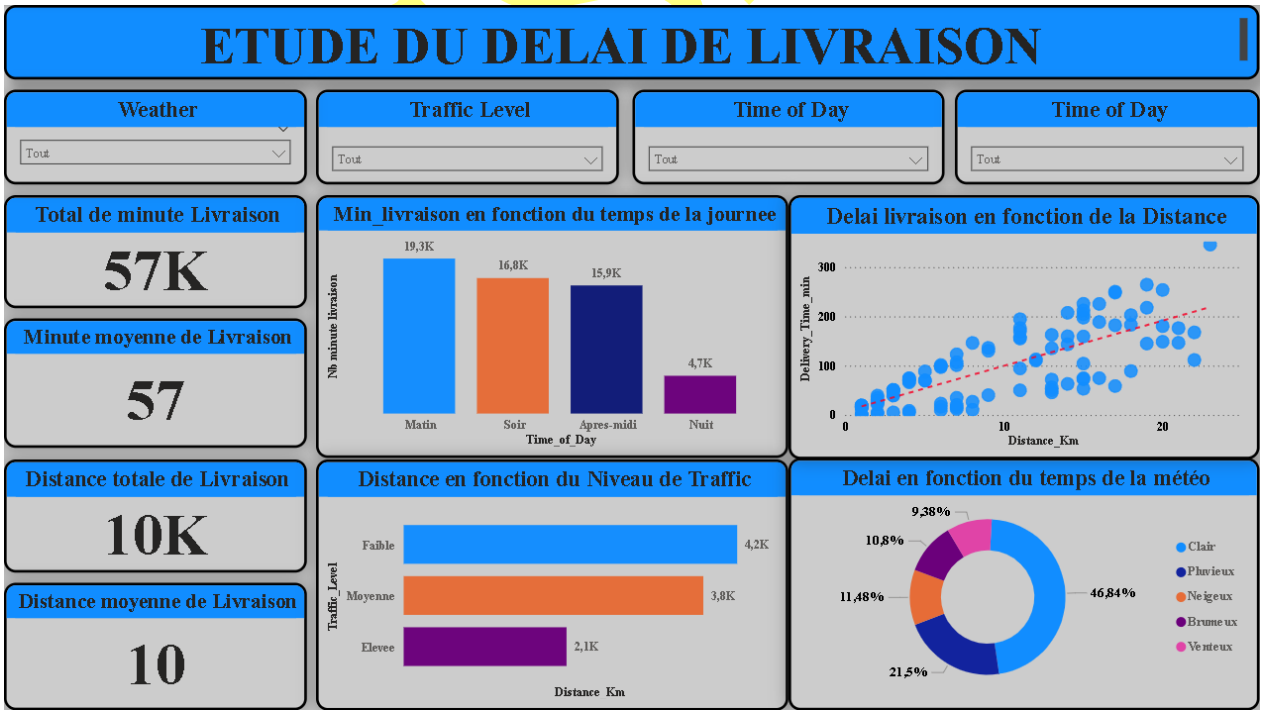
5.Communication avec les clients:

- **Informer les clients des retards potentiels lors de conditions météorologiques défavorables pour améliorer la satisfaction client.**
- **Mettre en place des systèmes de notification en temps réel pour tenir les clients informés de l'état de leur livraison.**

6.Collecte de données supplémentaires : - Enrichir le jeu de données avec des informations plus détaillées sur les conditions météorologiques (intensité de la neige, vitesse du vent, etc.) et d'autres facteurs contextuels. - Utiliser des données en temps réel pour améliorer la précision des prévisions de délais de livraison.

7.Validation croisée des résultats : - Tester le modèle sur un autre jeu de données pour valider la généralisable des résultats. - Utiliser des techniques de validation croisée pour évaluer la performance du modèle.

POWER-BI



CODE R et POWER-QUERY

- Transformation et Doublons

```
let

    Source                                                                    =
Csv.Document(File.Contents("C:\Users\HP\Downloads\INSSEDS\cours\ECONOMET
RIE\Mini       projet       économétrie       RegMult       et       ANO-
VA\delai_livraison.csv"),[Delimiter=";", Columns=9, Encoding=1252, Quo-
teStyle=QuoteStyle.None]),

    #"En-têtes promus" = Table.PromoteHeaders(Source, [PromoteAllSca-
lars=true]),

    #"Type modifié" = Table.TransformColumnTypes(#"En-têtes pro-
mus",{{"Order_ID", type text}, {"Distance_km", type text}, {"Weather",
type text}, {"Traffic_Level", type text}, {"Time_of_Day", type text},
{"Vehicle_Type", type text}, {"Preparation_Time_min", Int64.Type},
{"Courier_Experience_yrs", type text}, {"Delivery_Time_min",
Int64.Type}}),

    #"Valeur remplacée" = Table.ReplaceValue(#"Type modi-
fié",".","",",Replacer.ReplaceText,{"Distance_km"}),

    #"Type modifié1" = Table.TransformColumnTypes(#"Valeur rempla-
cée",{{"Distance_km", type number}}),

    #"Valeur remplacée1" = Table.ReplaceValue(#"Type modi-
fié1","Clear","Clair",Replacer.ReplaceText,{"Weather"}),

    #"Valeur remplacée2" = Table.ReplaceValue(#"Valeur rempla-
cée1","Rainy","Pluvieux",Replacer.ReplaceText,{"Weather"}),

    #"Valeur remplacée3" = Table.ReplaceValue(#"Valeur rempla-
cée2","Snowy","Neigeux",Replacer.ReplaceText,{"Weather"}),

    #"Valeur remplacée4" = Table.ReplaceValue(#"Valeur rempla-
cée3","Windy","Venteux",Replacer.ReplaceText,{"Weather"}),

    #"Valeur remplacée5" = Table.ReplaceValue(#"Valeur rempla-
cée4","Foggy","Brumeux",Replacer.ReplaceText,{"Weather"}),

    #"Valeur remplacée6" = Table.ReplaceValue(#"Valeur rempla-
cée5","Low","Faible",Replacer.ReplaceText,{"Traffic_Level"}),

    #"Valeur remplacée7" = Table.ReplaceValue(#"Valeur rempla-
cée6","Medium","Moyenne",Replacer.ReplaceText,{"Traffic_Level"}),

    #"Valeur remplacée8" = Table.ReplaceValue(#"Valeur rempla-
cée7","High","Elevee",Replacer.ReplaceText,{"Traffic_Level"}),

    #"Valeur remplacée9" = Table.ReplaceValue(#"Valeur rempla-
cée8","Morning","Matin",Replacer.ReplaceText,{"Time_of_Day"}),

    #"Valeur remplacée10" = Table.ReplaceValue(#"Valeur rempla-
cée9","Evening","Soir",Replacer.ReplaceText,{"Time_of_Day"}),
```

```
    #"Valeur remplacée11" = Table.ReplaceValue(#"Valeur rempla-
cée10","Night","Nuit",Replacer.ReplaceText,{"Time_of_Day"}),

    #"Valeur remplacée12" = Table.ReplaceValue(#"Valeur rempla-
cée11","Afternoon","Apres-midi",Replacer.ReplaceText,{"Time_of_Day"}),

    #"Valeur remplacée13" = Table.ReplaceValue(#"Valeur rempla-
cée12","Bike","Velo",Replacer.ReplaceText,{"Vehicle_Type"}),

    #"Valeur remplacée14" = Table.ReplaceValue(#"Valeur rempla-
cée13","Car","Voiture",Replacer.ReplaceText,{"Vehicle_Type"}),

    #"Valeur remplacée15" = Table.ReplaceValue(#"Valeur rempla-
cée14",".",",",Replacer.ReplaceText,{"Courier_Experience_yrs"}),

    #"Type modifié2" = Table.TransformColumnTypes(#"Valeur rempla-
cée15",{{"Courier_Experience_yrs", Int64.Type}}),

    #"Doublons supprimés" = Table.Distinct(#"Type modifié2", {"Or-
der_ID"})

in

    #"Doublons supprimés"
```

- **Pretraitrement des données**
- **Visualisations des donnees**

```
library(readxl)

library(dplyr)

Delai_livraison                                                    <-
read_excel("C:\\Users\\HP\\Downloads\\INSSEDS\\cours\\ECONOMETRIE\\Mini
projet économétrie RegMult et ANOVA\\Delai_livraison.xlsx")

Delai_livraison = Delai_livraison %>%

  select(Order_ID,Delivery_Time_min,everything())

str(Delai_livraison)
```

- **Traitement des valeurs manquantes**

```
Delai_livraison$Order_ID = as.character(Delai_livraison$Order_ID)

Delai_livraison$Weather  = as.factor(Delai_livraison$Weather)

Delai_livraison$Traffic_Level = as.factor(Delai_livraison$Traffic_Level)

Delai_livraison$Time_of_Day = as.factor(Delai_livraison$Time_of_Day)

Delai_livraison$Vehicle_Type = as.factor(Delai_livraison$Vehicle_Type)

Delai_livraison
```

- **Résumé des variables avant traitement des valeurs manquantes**

```
summary (Delai_livraison)
```

- **Résumé des variables après traitement des valeurs manquantes**

```
library(VIM)

Delai_livraison= kNN(Delai_livraison)

Delai_livraison                                                    =
subset(Delai_livraison,select=Order_ID:Courier_Experience_yrs)

summary(Delai_livraison)
```

- **Traitement des valeurs abberantes**
- **Visualisation des valeurs abberantes**

```
library(onesime)

onesime_boites_a_moustaches_avec_outliers(Delai_livraison)
```

- **Visualisation après traitement**

```
library(onesime)

library(DescTools)

De-
lai_livraison$Delivery_Time_min=Winsorize(Delai_livraison$Delivery_Time_
min)

onesime_boites_a_moustaches_avec_outliers(Delai_livraison)
```

- **ANALYSE UNIVARIEE**
- **ANALYSE DES VARIABLES QUANTITATIVES**
- **Variable "Delivery_Time_min"**
- **Tableau**

```
library(akposso)

head(akposso.qt.tableau(Delai_livraison$Delivery_Time_min),5)
```

- **Graphique**

```
akposso.qt.graph(Delai_livraison$Delivery_Time_min)
```

- **Resumés numériques**

```
akposso.qt.resume(Delai_livraison$Delivery_Time_min)
```

- **Variable "Distance_km"**
- **Tableau**

```
head(akposso.qt.tableau(Delai_livraison$Distance_km),5)
```

- **Graphique**

```
akposso.qt.graph(Delai_livraison$Distance_km)
```

- **Résumé numérique**

```
akposso.qt.resume(Delai_livraison$Distance_km)
```

- **ANALYSE DES VARIABLES QUALITATIVES**
- **Variable "Météo"**
- **Tableau**

```
akposso.ql.tableau(Delai_livraison$Weather)
```

- **Graphique**

```
akposso.ql.graph(Delai_livraison$Weather)
```

- **Variables "Traffic_Level"**
- **Tableau**

```
akposso.ql.tableau(Delai_livraison$Traffic_Level)
```

- **Graphique**

```
akposso.ql.graph(Delai_livraison$Traffic_Level)
```

- **Variable "Time_of_Day"**
- **Tableau**

```
akposso.ql.tableau(Delai_livraison$Time_of_Day)
```

- **Graphique**

```
akposso.ql.graph(Delai_livraison$Time_of_Day)
```

- **ANALYSE BIVARIEE**
- **Variables "Delivery_Time_min" et "Distance_Km"**
- **Graphique**

```
plot(Delai_livraison$Distance_km,Delai_livraison$Delivery_Time_min,main="Nuage de points", xlab="Distance_Km",ylab="Delivery_Time_min",pch = 20,col="blue")

abline(lm(Delai_livraison$Delivery_Time_min ~Delai_livraison$Distance_km), col="red", lwd = 2)
```

- **Equation de la droite**

```
lm(Delai_livraison$Delivery_Time_min ~ Delai_livraison$Distance_km)
```

- **Liaison**

```
library(akposso)

akpos-so.2qt.liaison(Delai_livraison$Delivery_Time_min,Delai_livraison$Distance_km)
```

- **Variables "Delivery_Time_min" et "Météo"**
- **Graphique**

```
boxplot(Delai_livraison$Delivery_Time_min ~ Delai_livraison$Weather, col=c("Orange","green"), xlab

= "Weather",ylab="Delivery_Time_min")
```

- **Liaison**

```
akpos-so.qtql.liaison(Delai_livraison$Delivery_Time_min,Delai_livraison$Weather)
```

- **Variables "Delivery_Time_min" et "Traffic_Level"**
- **Graphique**

```
boxplot(Delai_livraison$Delivery_Time_min~Delai_livraison$Traffic_Level, col=c("Orange","green"),xlab = "Traffic_Level",ylab="Delivery_Time_min")
```

- **Liaison**

```
akpos-so.qtql.liaison(Delai_livraison$Delivery_Time_min,Delai_livraison$Traffic_Level)
```

- **Variables "Delivery_Time_min" et "Time_of_Day"**
- **Graphique**

```
boxplot(Delai_livraison$Delivery_Time_min ~ Delai_livraison$Time_of_Day,
col=c("Orange","green"), xlab = "Time_of_Day",ylab="Delivery_Time_min")
```

- **Liaison**

```
akposso.qtql.liaison(Delai_livraison$Delivery_Time_min, De-
lai_livraison$Time_of_Day)
```

- **ESTIMATION DE LA MOYENNE DE LA VARIABLE CIBLE "Delivery_Time_min"**
- **Tester si la variable suit la loi normale**

```
shapiro.test(Delai_livraison$Delivery_Time_min)
```

- **Estimation**

```
library(RVAideMemoire)

boot-
strap(Delai_livraison$Delivery_Time_min,function(x,i)mean(Delai_livraiso
n$Delivery_Time_min[i]))
```

- **TEST DE CONFORMITE DE LA VARIABLE CIBLE**
- **Conditions**

```
shapiro.test(Delai_livraison$Delivery_Time_min)
```

- **Test**

```
wilcox.test(Delai_livraison$Delivery_Time_min,mu=57)
```

- **TEST DE LIAISONS**
- **variable "Delivery_Time" et "Distance_Km"**
- **Normalité des variable "Delivery_Time" et "Distance_Km"**

```
shapiro.test(Delai_livraison$Delivery_Time_min)

shapiro.test(Delai_livraison$Distance_km)
```

- **Test de liaison**

```
library(RVAideMemoire)

cor.test(Delai_livraison$Delivery_Time_min,Delai_livraison$Distance_km,m
ethod="kendall")

cor.test(Delai_livraison$Delivery_Time_min,Delai_livraison$Distance_km,m
ethod = "kendall",alternative = "greater")
```

- **Variable "Delivery_Time" et "Weather"**
- **1ere étape : comparer graphiquement les deux sous population**

```
boxplot(Delai_livraison$Delivery_Time_min~Delai_livraison$Weather,xlab =
"Weather",ylab = "Delivery_Time_min")
```

- **2eme étape : tester la normalité des données dans chaque sous population**

```
library(RVAideMemoire)

byf.shapiro(Delai_livraison$Delivery_Time_min~Delai_livraison$Weather)
```

- **3eme étape : tester l'égalité des variances**

```
library(car)

leveneTest(Delai_livraison$Delivery_Time_min~Delai_livraison$Weather)
```

- **4eme étape: test de liaison**

```
kruskal.test(Delivery_Time_min ~ Weather, data = Delai_livraison)
```

- **variable "Delivery_Time" et "Traffic_level"**
- **1ere étape : comparer graphiquement les deux sous population**

```
boxplot(Delai_livraison$Delivery_Time_min ~ De-
lai_livraison$Traffic_Level,xlab = "Traffic_level",ylab = "Deliv-
ery_Time_min")
```

- **2eme étape : tester la normalité des données dans chaque sous population**

```
library(RVAideMemoire)

byf.shapiro(Delai_livraison$Delivery_Time_min~Delai_livraison$Traffic_Le
vel)
```

- **3eme étape : tester l'égalité des variances**

```
library(car)

le-
ve-
neTest(Delai_livraison$Delivery_Time_min~Delai_livraison$Traffic_Level)
```

- **4eme étape : Test de liaison**

```
kruskal.test(Delivery_Time_min ~ Traffic_Level, data = Delai_livraison)
```

- **Variable "Delivery_Time" et "Time_of_day"**

- **1ere étape : comparer graphiquement les deux sous population**

```
boxplot(Delai_livraison$Delivery_Time_min ~ De-
lai_livraison$Time_of_Day,xlab = "Time_of_day",ylab
="Delivery_Time_min")
```

- **2eme étape : tester la normalité des données dans chaque sous population**

```
library(RVAideMemoire)

byf.shapiro(Delai_livraison$Delivery_Time_min~Delai_livraison$Time_of_Da
y)
```

- **3eme étape : tester l'égalité des variances**

```
library(car)

leveneTest(Delai_livraison$Delivery_Time_min~Delai_livraison$Time_of_Day
)
```

- **4eme étape: test de liaison**

```
kruskal.test(Delivery_Time_min ~ Time_of_Day, data = Delai_livraison)
```

- **MODELISATION REGRESSION MULTIPLE**

- **1ere étape: Importer les données**

```
library(dplyr)

df = Delai_livraison %>%

  se-
lect(Delivery_Time_min,Distance_km,Weather,Traffic_Level,Time_of_Day)

str(df)
```

- **2eme étape : Estimer les paramètres**
- **Faire la regression**

```
regM <- lm(df$Delivery_Time_min ~ df$Distance_km + df$Weather +
df$Traffic_Level + df$Time_of_Day)

regM$coefficients
```

- **Voir l'intervalle de confiance des coefficients estimés**

```
confint(regM)
```

- **3eme étape : Test de signifucativité Globale**

```
summary(regM)
```

- **4eme étape : Test de significativité individuel**

```
print(anova(regM,test="Chisq"))

summary(regM)
```

- **5eme étape : Analyse des résidus**

- Graphique

```
res.m<-rstudent(regM)

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

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

res.m<-rstudent(regM)

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

- **6eme étape : Analyse de la multicolinéarité**
- **modele avec la variable Time_of_Day**

```
# Charger les bibliothèques nécessaires

library(stats)


# Supposons que vous avez déjà estimé un modèle avec lm()

# Exemple : modèle <- lm(Y ~ X1 + X2 + X3, data = votre_data)


# Fonction pour extraire les indicateurs

evaluer_modele <- function(regM) {
```

```
  # Récupérer le R² ajusté
  r2_ajuste <- summary(regM)$adj.r.squared

  # Récupérer l'AIC
  aic <- AIC(regM)

  # Récupérer le BIC
  bic <- BIC(regM)

  # Récupérer le test F global
  test_f <- summary(regM)$fstatistic
  p_value_f <- pf(test_f[1], test_f[2], test_f[3], lower.tail = FALSE)

  # Afficher les résultats
  cat("R² ajusté :", r2_ajuste, "\n")
  cat("AIC :", aic, "\n")
  cat("BIC :", bic, "\n")
  cat("Test F (valeur, df1, df2) :", test_f, "\n")
  cat("p-value du test F :", p_value_f, "\n")
}

# Appliquer la fonction à votre modèle
evaluer_modele(regM)
```

- **modele sans la variable Time of day**

```
regB <- lm(df$Delivery_Time_min ~ df$Distance_km + df$Weather +
df$Traffic_Level)
# Charger les bibliothèques nécessaires
library(stats)

# Supposons que vous avez déjà estimé un modèle avec lm()
# Exemple : modèle <- lm(Y ~ X1 + X2 + X3, data = votre_data)
```

```
# Fonction pour extraire les indicateurs
evaluer_modele <- function(regB) {
  # Récupérer le R² ajusté
  r2_ajuste <- summary(regB)$adj.r.squared

  # Récupérer l'AIC
  aic <- AIC(regB)

  # Récupérer le BIC
  bic <- BIC(regB)

  # Récupérer le test F global
  test_f <- summary(regB)$fstatistic
  p_value_f <- pf(test_f[1], test_f[2], test_f[3], lower.tail = FALSE)

  # Afficher les résultats
  cat("R² ajusté :", r2_ajuste, "\n")
  cat("AIC :", aic, "\n")
  cat("BIC :", bic, "\n")
  cat("Test F (valeur, df1, df2) :", test_f, "\n")
  cat("p-value du test F :", p_value_f, "\n")
}

# Appliquer la fonction à votre modèle
evaluer_modele(regB)
```

- **Multicolinéarité**

```
library(olsrr)
ols_vif_tol(regB)
```

- **7eme étape: Validation du modèle**
- **Test de linéarité du modèle**

```
library(lmtest)
```

```
raintest(regB)
```

- **regression avec interaction**

```
regC = regB <- lm(df$Delivery_Time_min ~ df$Distance_km + df$Weather +
df$Traffic_Level + df$Weather * df$Traffic_Level)

raintest(regC)

library(olsrr)

ols_vif_tol(regC)
```

- **Homoscedasticité des erreurs**
- **Graphique**

```
plot(regB,1)
```

- **TEST BREUSCH-PAGAN**

```
library(lmtest)

bptest(regB)
```

- **Autocorrélation des erreurs**
- **Graphique**

```
acf(residuals(regB))
```

- **TEST DE DURBIN-WATSON**

```
library(lmtest)

dwtest(regB)
```

- **Normalité des erreurs**

```
shapiro.test(residuals(regB))
```

- **MODELISATION ANOVA-1**
- **1ere étape : comparer graphiquement les sous populations**

```
box-
plot(Delai_livraison$Delivery_Time_min~Delai_livraison$Weather,xlab="Wea
ther",ylab="Delivery_Time_min")
```

- **2eme étape : estimer les statistiques de base (mean, quantile, sd) par ss pop**
- **Moyenne(mean)**

```
tap-
ply(Delai_livraison$Delivery_Time_min,Delai_livraison$Weather,mean,na.rm
=TRUE)
```

- **Quantile**

```
tap-
ply(Delai_livraison$Delivery_Time_min,Delai_livraison$Weather,quantile,n
a.rm=TRUE)
```

- **Ecart-type(sd)**

```
tap-
ply(Delai_livraison$Delivery_Time_min,Delai_livraison$Weather,sd,na.rm=T
RUE)
```

- **3eme étape : tester la normalité des données dans chaque sous population**
- **Graphique**

```
library(car)

library(RVAideMemoire)

byf.qqnorm(Delai_livraison$Delivery_Time_min~Delai_livraison$Weather)
```

- **test de normalité**

```
library(RVAideMemoire)

byf.shapiro(Delai_livraison$Delivery_Time_min~Delai_livraison$Weather)
```

- **4eme étape : tester l'égalité des variances**

```
kruskal.test(Delai_livraison$Delivery_Time_min~Delai_livraison$Weather)
```

- **5eme étape : faire un test robuste par bootsrap (réé-chantillonage)**

```
library(pgirmess)

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

PermTest(reg.aov, B=1000)
```

- **6eme étape : tester la significativité du facteur: tester l'égalité des moyennes**

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

anova(reg.aov)

library(RVAideMemoire)

pair-
wise.perm.t.test(Delai_livraison$Delivery_Time_min,Delai_livraison$Weath
er)
```

- **6eme étape : Analyser les résidus**

```
library(lattice)

res.aov<-rstudent(reg.aov)

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

res.aov<-rstudent(reg.aov)

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

- **7eme étape: Interpreter les coefficients**

```
summary(reg.aov)
```

- **Si on prend la contrainte ($\sum \alpha i=0$) ce qui revient à prendre la moyenne comme référence**.

```
summary(lm(Delivery_Time_min~C(Weather,sum),data=Delai_livraison))
```

- **Si on fait le test de significativité**

```
# Extraire la matrice de variance-covariance des coefficients

vcov_matrix <- vcov(reg.aov)


# Calculer la variance de alpha_5

var_alpha5 <- sum(vcov_matrix[1:4, 1:4]) + 2 * sum(vcov_matrix[1, 2:4])
+ 2 * sum(vcov_matrix[2, 3:4]) + 2 * vcov_matrix[3, 4]


# Calculer l'erreur standard de alpha_5

se_alpha5 <- sqrt(var_alpha5)


# Calculer la statistique t
```

```
alpha5 <- -3.3645  # Valeur de alpha_5 calculée
t_alpha5 <- alpha5 / se_alpha5

# Calculer la p-value
p_value_alpha5 <- 2 * pt(-abs(t_alpha5), df = df.residual(reg.aov))

# Afficher les résultats
cat("alpha5:", alpha5, "\n")
cat("SE(alpha5):", se_alpha5, "\n")
cat("t-value:", t_alpha5, "\n")
cat("p-value:", p_value_alpha5, "\n")
```