HEC MONTRÉAL

Devoir 1

Travail présenté à

Linda Mhalla

Dans le cadre du cours

Analyse et inférence statistique

MATH60619.H2020

Par

Carl Martel

11127417

Adil Labiad

11279549

19 février 2020

Devoir 1

Question 1.

Analyse exploratoire des données *Renfe*

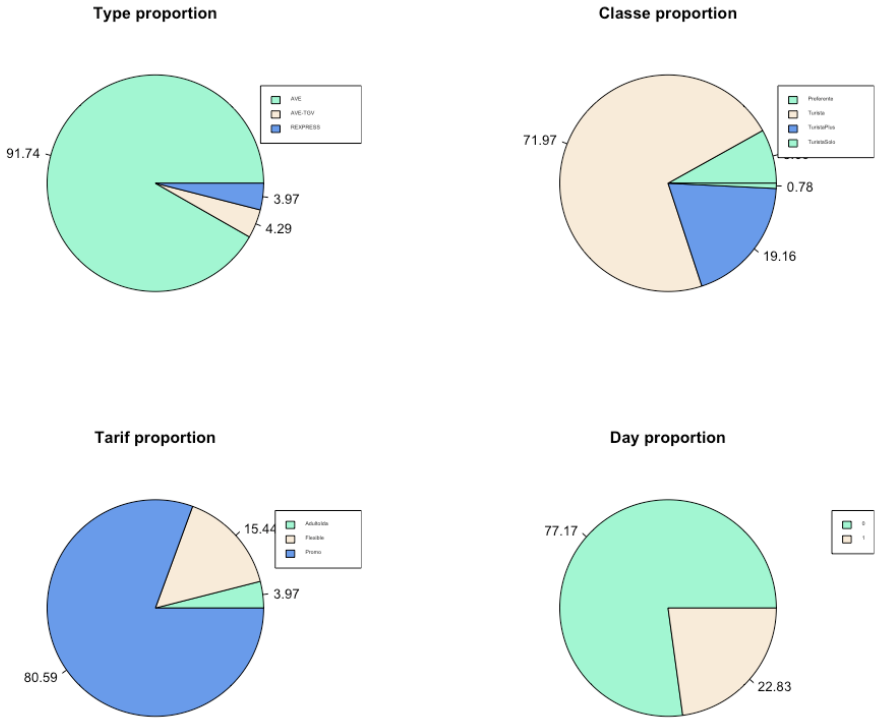
Tableau 1 : Statistiques descriptives

```
      prix                type               classe          tarif
 Min.   : 32.3   AVE     :9174   Preferente : 809   AdultoIda: 397
 1st Qu.: 75.4   AVE-TGV : 429   Turista    :7197   Flexible :1544
 Median : 85.1   REXPRESS: 397   TuristaPlus:1916   Promo    :8059
 Mean   : 86.1                   TuristaSolo:  78
 3rd Qu.:100.4
 Max.   :214.2
      dest              duree            jour
 Min.   :0.0000   Min.   :150.0   Min.   :1.000
 1st Qu.:0.0000   1st Qu.:150.0   1st Qu.:2.000
 Median :1.0000   Median :170.0   Median :4.000
 Mean   :0.5062   Mean   :185.8   Mean   :3.917
 3rd Qu.:1.0000   3rd Qu.:190.0   3rd Qu.:6.000
 Max.   :1.0000   Max.   :562.0   Max.   :7.000
```

Tableau 2 : Partie du tableau de fréquence type de train et prix

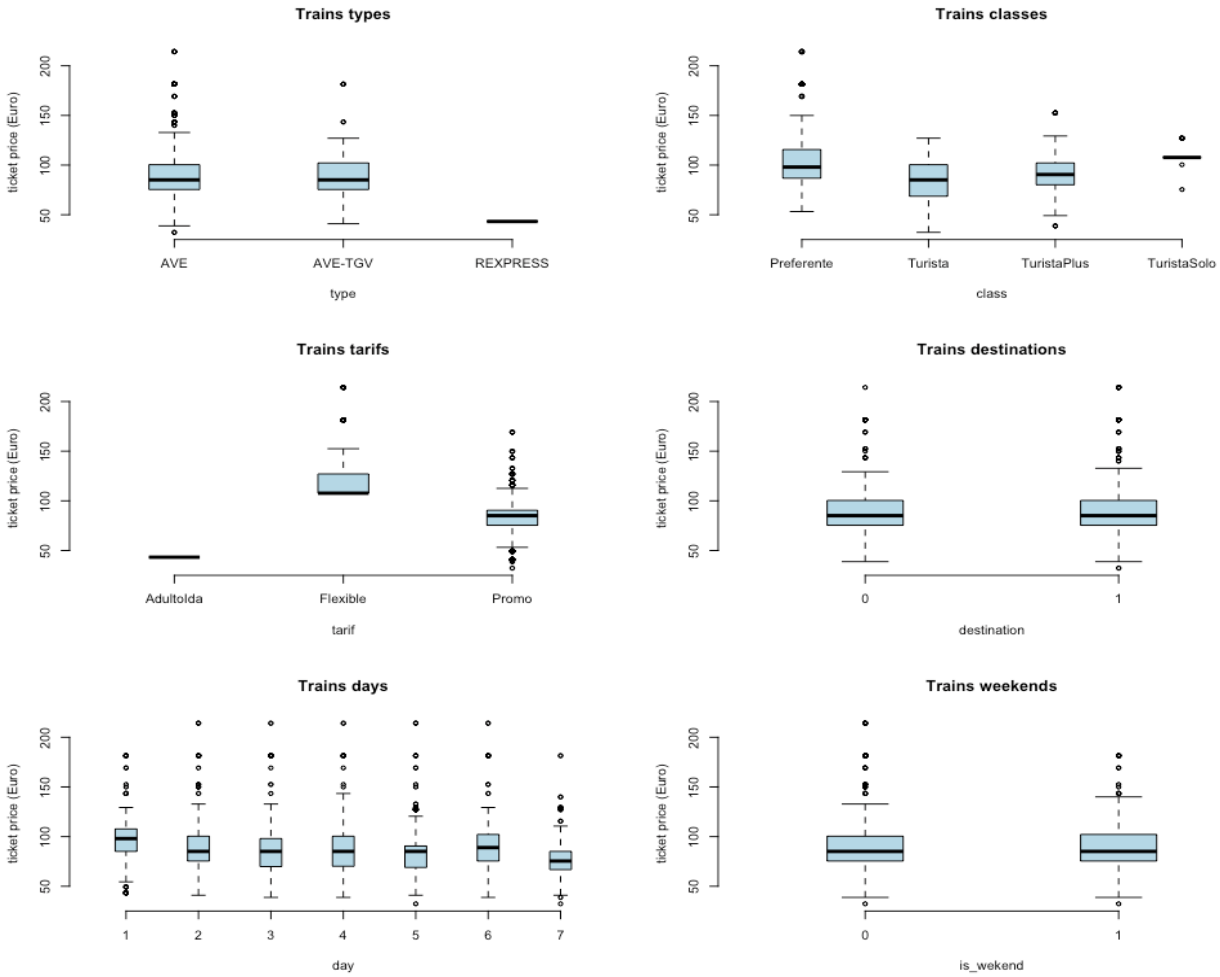
	AVE	AVE-TGV	REXPRESS
32.3	2	0	0
38.8	4	0	0
40.95	45	1	0
43.25	0	0	397

Figure 1 : Pie charts



La **figure 1** nous montre les différentes proportions des différentes classes. 91.74% des types de trains utilisés dans l'échantillon sont de type *AVE*. *Turista* est la classe la plus utilisée (71.97%). Au niveau du tarif, 80.59% sont des billets *Promo*. Enfin. 77.17% des gens ont pris le train la semaine, versus 22.83% durant le week-end.

Figure 2 : Boîtes à moustaches prix en relation avec les autres variables



La **figure 2** montre la relation entre le prix d'un billet de train de la compagnie *Renfe* et les différentes variables du cas.

Le jour de la semaine semble avoir une incidence sur le prix. Le prix moyen du samedi est le plus bas et celui du dimanche est le plus élevé.

Le prix varie également en fonction du type de train. Les prix moyens ne semblent pas varier entre les trains grande vitesse *AVE* et *AVE-TGV*. Cependant, les prix sont différents pour le train *REXPRESS*. Ce dernier est le moins dispendieux en moyenne; c'est en fait le train régional abordable et moins rapide. 397 personnes ont utilisé ce type de train (**tableau 1**) et le prix est fixé à 43.25€ (**tableau 2**).

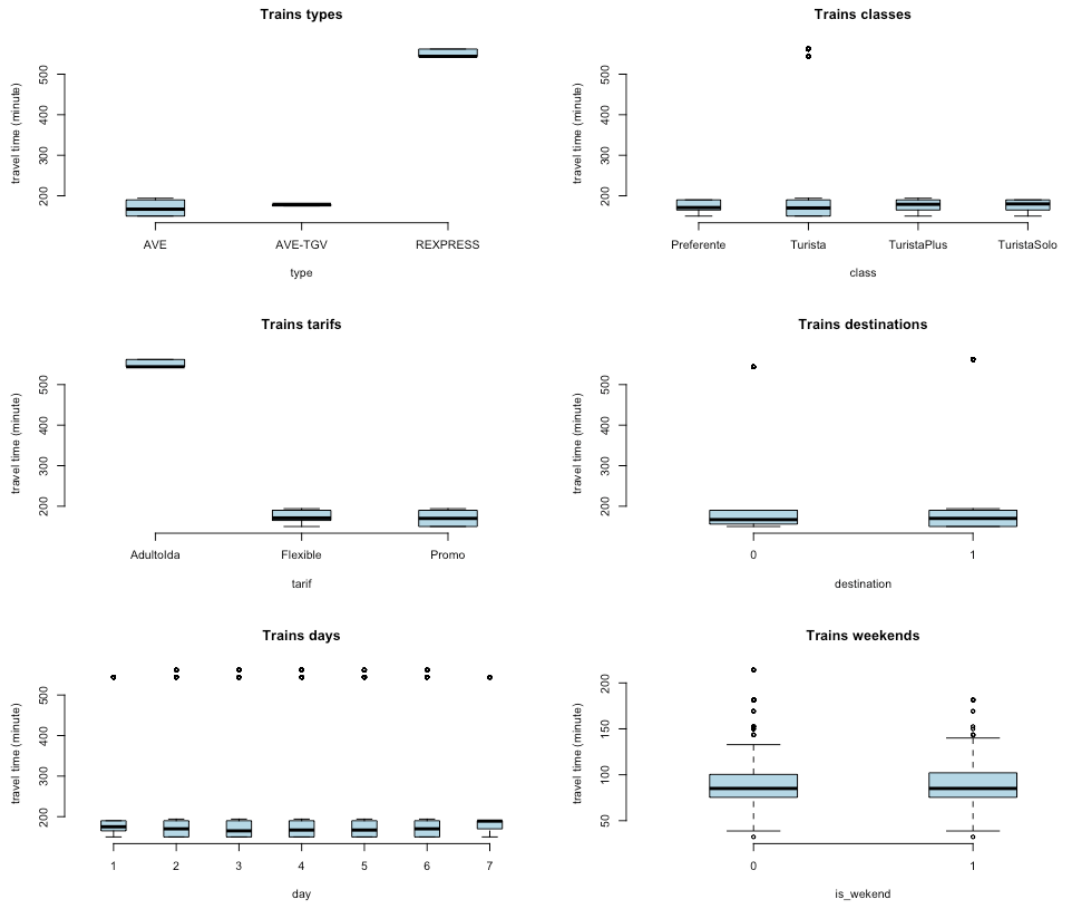
Les tarifs impactent aussi le prix. Selon la **figure 2**, les prix du tarif *Flexible* sont plus élevés. Selon le site web de *Renfe*, le tarif *Flexible* inclut des options additionnelles permettant au passager d'obtenir de meilleures conditions pour faire des changements, pour annuler ou encore pour le cas où il manque le train.[1] Le tarif *Promo* est de base, sans avantages supplémentaires.[2] Le

[1] http://www.renfe.com/EN/viajeros/tarifas/billete_flexible.html
[2] http://www.renfe.com/EN/viajeros/tarifas/billete_promo.html

tarif *AdultoIda* est le moins dispendieux et a été toujours constant à un prix de 43.25€, le même que celui du train de type *REXPRESS*. Selon le **tableau 1**, 397 personnes ont choisi ce tarif et ce même nombre ont utilisé le train *REXPRESS*; il est donc possible de déduire que ce type de tarif est celui utilisé pour la réservation dans les trains régionaux.

Ensuite, la classe du billet a aussi un impact sur le prix. La classe *Preferente* est la plus chère en moyenne; c'est la classe supérieure. Il n'y a pas beaucoup d'observations, mais la classe *TuristaSolo* semble être aussi dispendieuse que la classe *Preferente*. La classe *Turista* est standard et la classe *TuristaPlus* ajoute un peu d'avantages.

Figure 3 : Relation entre durée et type de train et relation entre la durée et le tarif



Au niveau de la durée, le trajet dans un train grande vitesse va prendre moins de temps que dans un train régulier (**figure 3**). La durée élevée du trajet avec un billet de tarif *AdultoIda* est expliquée par le fait que ce type de billet est utilisé pour la réservation des trains *REXPRESS*.

Question 2.

Tableau 3 : Statistiques descriptives pour n=1000

```
#       prix             type           classe         tarif         dest       duree
# Min.   : 40.95   AVE     :921   Preferente :420   AdultoIda: 29   Min.   :0   Min.   :150.0
# 1st Qu.: 75.40   AVE-TGV : 50   Turista    :580   Flexible :147   1st Qu.:0   1st Qu.:165.0
# Median : 88.95   REXPRESS: 29   TuristaPlus:  0   Promo    :824   Median :0   Median :170.0
# Mean   : 92.02                  TuristaSolo:  0                   Mean   :0   Mean   :182.9
# 3rd Qu.:100.40                                                    3rd Qu.:0   3rd Qu.:190.0
# Max.   :214.20                                                    Max.   :0   Max.   :544.0
```

Ce ne serait pas un bon échantillon. Comme on peut voir dans le **tableau 3**, sous la colonne *dest*, cet échantillon ne contient que des données dont le trajet est de Barcelone vers Madrid. Ce n'est clairement pas un échantillon aléatoire. Aussi, on voit que dans le nouvel échantillon, le nombre d'observations dans la classe *TuristaPlus* et *TuristaSolo* est nul, contrairement à notre échantillon de base qui contenait 19% des gens pour *TuristaPlus* et 0.78% pour la classe *TuristaSolo* (**tableau 1**). Ce n'est donc pas représentatif de l'échantillon initial.

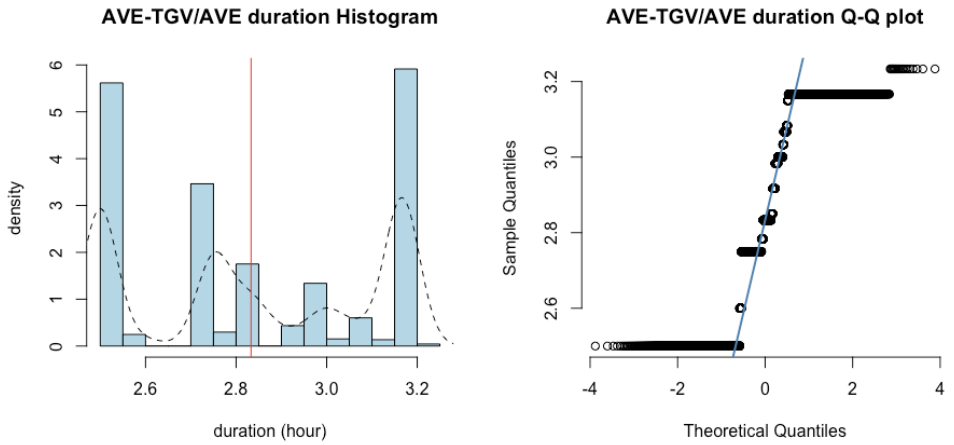
Question 3.

a) À mesure que l'on a une valeur de μ_0 ou ν_0 qui s'éloigne de la vraie valeur de la moyenne ou médiane dans la population pour les trains à grande vitesse, plus on a de chance de rejeter l'hypothèse nulle H_o : $\mu=\mu_0$ ou H_0 : $\nu=\nu_0$, car μ_0 et ν_0 diffère de plus en plus de la vraie valeur.

b) Toute autre chose étant égale par ailleurs, si on augmente la taille de l'échantillon, les tests d'hypothèses seront plus efficaces à rejeter l'hypothèse nulle quand elle est fausse. Lorsqu'on augmente la taille de l'échantillon, l'erreur-type diminue, ce qui fait diminuer l'intervalle de confiance. Par exemple, pour l'hypothèse H_o : $\mu_0=2.83$, un intervalle de confiance plus petit augmente les chances que 2.83 soit à l'extérieur de l'intervalle, donc plus grand taux de rejet de H_0. La puissance des tests va alors augmenter, ainsi que le pourcentage de valeurs-p inférieures à 0.05 (rejet de H_0). Les nouveaux points seront donc au-dessus de la courbe actuelle.

c) Lorsque $\mu=\mu_0$ ou $\nu=\nu_0$, les graphiques de loi normales de l'hypothèse nulle et de l'hypothèse alternative se superposent. La probabilité de rejet correspond donc à $\alpha=0.05$. Cela correspond à la probabilité de faire une erreur de type 1, c'est-à-dire à la probabilité de rejeter H_0 lorsque H_0 est vraie.

Figure 4 : Histogramme et Q-Q plot de la durée en heure des trains grande vitesse



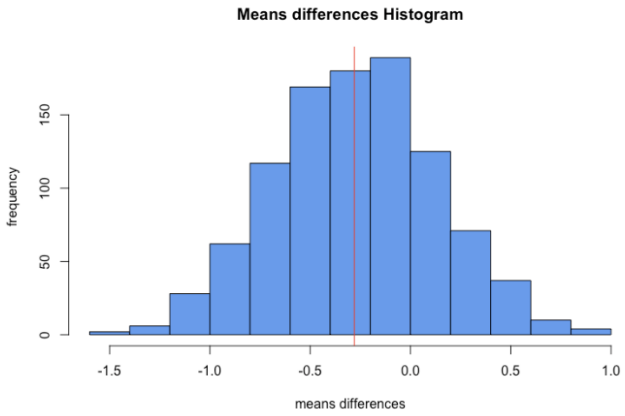
d) On rejetterait l'hypothèse nulle à $\nu=2.833$ à une fréquence d'environ 45%. L'histogramme de la **figure 4** montre que les données n'obéissent pas à une loi symétrique dans l'ensemble, excepté aux deux bouts du graphe. Beaucoup de données sont concentrées loin de la médiane, sur les extrêmes de l'histogramme. Cela peut expliquer le haut taux de rejet de l'hypothèse nulle du test de Wilcoxon au niveau $\nu=\nu_0$.

e) Le postulat de normalité du test-t n'est pas valide. Dans le *Q-Q plot* de la **figure 4**, la plupart des points s'éloignent de la droite normale. La robustesse du test-t disparaît quand la loi normale n'est pas respectée.

Question 4.

a) Le taux de couverture empirique des intervalles de confiances à 95% est de 94.7%.

b)



c) La puissance du test est de 10.5%

Question 5.

Hypothèses

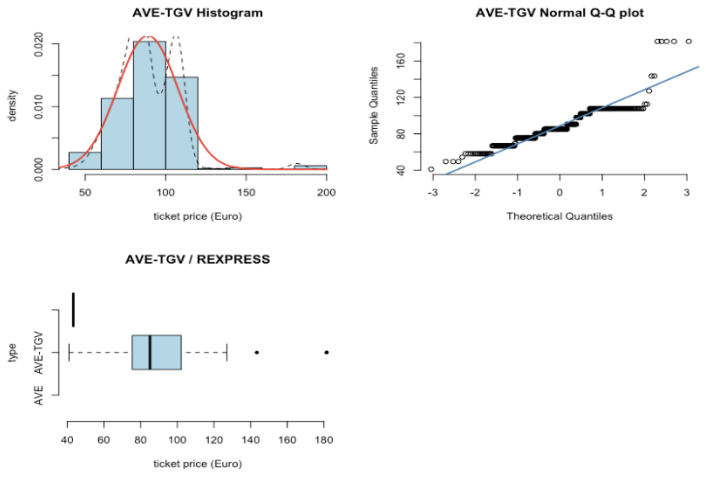
H_0: $\mu_{AVE\text{-}TGV} = (\mu_{REXPRESS}=43.25€)$ H_1: $\mu_{AVE\text{-}TGV} \neq (\mu_{REXPRESS}=43.25€)$

Où $\mu_{AVE\text{-}TGV}$ est le prix moyen en euros du billet pour un train de type *AVE-TGV,*

$\mu_{REXPRESS}$ est le prix moyen en euros du billet pour un train de type *REXPRESS.*

Justification du choix de test :

Figure 5 : Graphiques pour le postulat de normalité



Au niveau des conditions de validités du test-t, l'indépendance des observations est respectée. On voit dans le graphique *Q-Q plot* à la **figure 5** que la normalité est plus ou moins respectée.

Nous avons choisi de faire un test-t pour un échantillon, car on compare le prix moyen du type de train *AVE-TGV* à celui du type *REXPRESS* qui lui est toujours constant à 43.25€. *REXPRESS* n'est pas une variable aléatoire.

La différence moyenne estimée entre les prix des deux types de train est de (88.88-43.25) = 45.63€. L'intervalle de confiance à 90% de cette différence : [44.14€; 47.12€] (Nous avons soustrait 43.25 aux deux bornes de l'intervalle de confiance généré dans R).

La valeur-p du test bilatéral du test-t pour un échantillon simple est inférieure à $2.2e^{-16}$. On rejette l'hypothèse nulle, car cette valeur est inférieure à α=0.1. Donc, le prix moyen du billet pour un train de type *AVE-TGV* est significativement différent que celui de *REXPRESS* au niveau de confiance 90%.

Question 6.

Hypothèses :

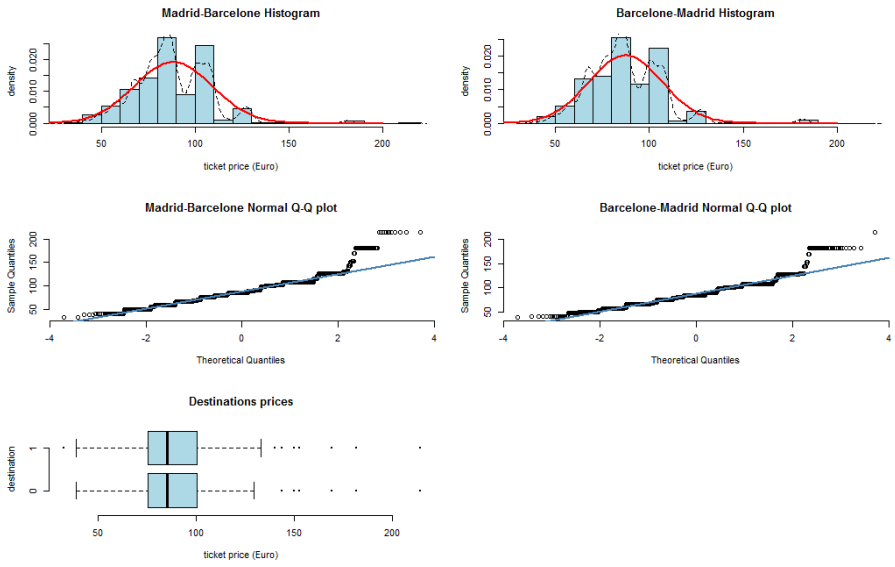
$H_0 : \mu_0 = \mu_1$ $H_1 : \mu_0 \neq \mu_1$

Où μ_0 est le prix moyen en euros pour les trains à grande vitesse sortants (Barcelone à Madrid)

μ_1 est le prix moyen en euros pour les trains à grande vitesse entrants (Madrid à Barcelone)

Test d'hypothèse :

Figure 6 : Graphiques postulat de normalité



Les données d'après la **figure 6** semblent suivre une loi normale selon les *Q-Q plots*. Cependant, vers la fin du graphique, les observations sont au-dessus de la droite. Le *boxplot* montre des valeurs extrêmes.

Nous avons testé l'égalité des variances, comme la valeur-p (0.00023) du test F est inférieure à α=0.05, on peut conclure qu'il y a une différence significative entre les variances des deux échantillons. On aura alors recours à un test de Welch. Nous ferons aussi le test non-paramétrique de Fligner-Policello, car le postulat de l'égalité des variances n'est pas nécessaire pour la validité de ce test.

La valeur-p (0.02156) du test de Welch bilatéral est inférieure à α=0.05, donc on rejette H_0.

La valeur-p (0.01181) du test de Fligner-Policello bilatéral est inférieure à α=0.05, donc on rejette H_0.

Il y a donc une différence significative entre le prix moyen des trains à grande vitesse sortants et le prix moyen des trains à grande vitesse entrants au niveau de confiance 95% autant pour le test paramétrique que non-paramétrique. Une direction est alors plus chère que l'autre.

Question 7.

Hypothèses :

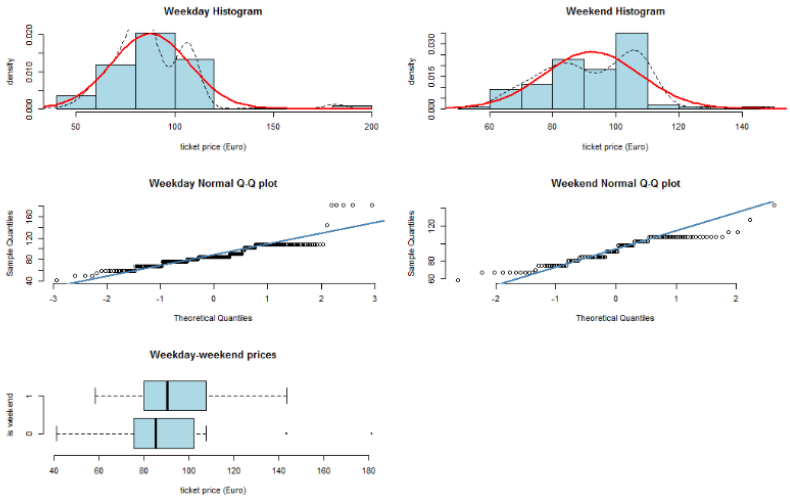
$H_0 : \mu_1 - \mu_0 \leq 0$ $H_1 : \mu_1 - \mu_0 > 0$

Où μ_0 est le prix moyen en euros des trains de type *AVE-TGV* un jour de semaine,

μ_1 est le prix moyen en euros des trains de type *AVE-TGV* un jour de fin de semaine.

Test d'hypothèses

Figure 7 : Graphiques postulat de normalité



Au niveau du postulat de normalité, les données semblent suivre une loi normale si l'on se fie au *Q-Q plots* précédents. Cependant, au bout des *Q-Q plots*, les données s'éloignent de la droite. Il y a quelques valeurs extrêmes selon le boxplot.

Nous avons testé l'égalité des variances. La p-value pour le test F est de 0.0016, inférieure à α=0.05. On rejette donc l'hypothèse nulle d'égalité des variances. On peut conclure qu'il y a une différence significative entre les variances des deux échantillons.

Nous avons décidé de faire un test non-paramétrique de Fligner-Policello qui sera robuste aux valeurs extrêmes et qui n'a pas besoin du postulat d'homoscédasticité pour être valide. La valeur-p du test de Fligner-Policello unilatéral est de 0.0006 et est inférieure à α=0.05, donc on rejette H_0.

Le prix moyen en euros d'un billet de train de type *AVE-TGV* la fin de semaine est plus élevé que le prix moyen des trains de type *AVE-TGV* un jour de semaine au niveau de confiance 95%.

Question 8.

a) $Prix = \beta_0 + \beta_1 dest + \beta_2 classe1 + \beta_3 classe2 + \beta_4 classe3 + \beta_5 duree + \beta_6 is\_weekend + \epsilon$

Où

- *Prix* est une variable continue correspondant au prix des billets de tarif *Promo* pour les trains de type *AVE* et *AVE-TGV*;
- *Dest* est une variable binaire du trajet, soit de Barcelone vers Madrid (0), ou de Madrid vers Barcelone (1);
- *Classe1=Turista, classe2=TuristaPlus, classe3=TuristaSolo*
  La catégorie de référence pour la variable *classe* est *Preferente;*
- *duree* est le temps du trajet annoncé en minutes;
- *is_weekend* est un indicateur binaire de la journée, soit un jour de la semaine (0) ou un jour de fin de semaine (1);

b) $\widehat{Prix} = 135.97 + 0.46 dest - 17.59 classe1 - 6.75 classe2 - 8.59 classe3 - 0.24 duree + 1.08 is\_weekend$

β_0: Lorsque toutes les co-variables sont nulles, le prix d'un billet de tarif *Promo* pour les trains grande vitesse est de 135.97€ en moyenne. L'interprétation n'est pas valide dans ce cas, car la variable *duree* ne peut être 0.
β_1: Toutes choses étant égales par ailleurs, pour le trajet de Madrid vers Barcelone, le prix d'un billet de tarif *Promo* pour les trains grande vitesse est de 0.46€ plus cher que pour le trajet Barcelone vers Madrid, en moyenne.
β_2: Toutes choses étant égales par ailleurs, le prix d'un billet de tarif *Promo* pour les trains grande vitesse de classe *Turista* est en moyenne 17.59€ moins cher qu'un billet de classe *Preferente*.

β_3: Toutes choses étant égales par ailleurs, le prix d'un billet de tarif *Promo* pour les trains grande vitesse de classe *TuristaPlus* est en moyenne 6.75€ moins cher qu'un billet de classe *Preferente*.

β_4: Toutes choses étant égales par ailleurs, le prix d'un billet de tarif *Promo* pour les trains grande vitesse de classe *TuristaSolo* est en moyenne 8.59€ moins cher qu'un billet de classe *Preferente*.

β_5: Toutes choses étant égales par ailleurs, pour une augmentation d'une minute de la durée annoncée du trajet, le prix d'un billet de tarif *Promo* pour les trains grande vitesse diminue en moyenne de 0.24€.

β_6: Toutes choses étant égales par ailleurs, le prix d'un billet de tarif *Promo* pour les trains grande vitesse est en moyenne 1.08€ plus cher la fin de semaine que la semaine.

c) *Dest :*

$H_0 : \beta_1=0$ $H_1 : \beta_1\neq0$

La p-valeur du test-t est de 0.127 et est supérieure à 0.05. On ne rejette pas H_0 à un niveau de test de 5%. La variable *dest* n'a donc pas d'effet significatif sur le prix.

Classe :

$H_0 : \beta_2=\beta_3=\beta_4=0$

H_1 : Au moins un des paramètres de la variable *classe* est utile pour prédire le prix et son effet linéaire est non-nul

La p-valeur du test F est inférieure à $2.2e^{-16}$. Cette valeur est inférieure à 0.05. On rejette H_0 à un niveau de test de 5%. Donc, au moins un des paramètres de la variable *classe* est utile pour prédire le prix et son effet linéaire est non-nul.

Duree :

$H_0 : \beta_5=0$ $H_1 : \beta_5\neq0$

La p-valeur du test-t est inférieure à $2.2e^{-16}$. Cette valeur est inférieure à 0.05. On rejette H_0 au niveau de test de 5%. La variable *duree* a donc un effet significatif sur le prix.

Is_weekend :

$H_0 : \beta_6=0$ $H_1 : \beta_6\neq0$

La p-valeur du test-t est de 0.0031 et est inférieure à 0.05. On rejette H_0 à un niveau de test de 5%. La variable *is_weekend* a donc un effet significatif sur le prix.

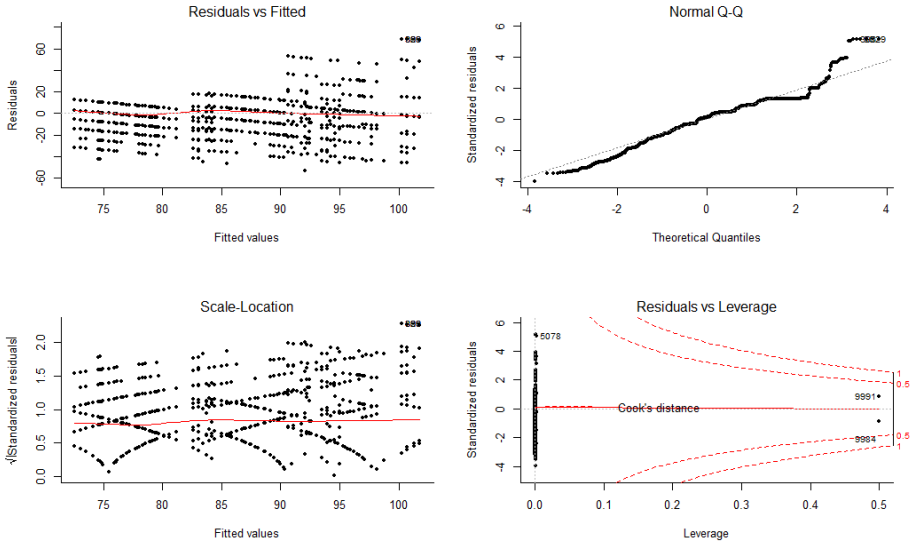
Modèle global :

$H_0 : \beta_1=\beta_2=\beta_3=\beta_4=\beta_5=\beta_6=0$

H_1 : Au moins une des variables du modèle est utile pour prédire le prix et son effet linéaire est non-nul

La valeur-p du test F est inférieure à $2.2e^{-16}$. Cette valeur est plus petite que 0.05. On rejette donc H_0 à un niveau de test de 5%. Au moins une des variables du modèle est utile pour prédire le prix et son effet linéaire est non-nul.

d) *Figure 8 : Diagnostic des résidus*



Linéarité:

Dans le graphique en haut à gauche de la **figure 8**, la droite de régression des résidus contre les valeurs ajustées est autour de 0. Les résidus ne suivent pas une forme quadratique ou autre. Ils sont bien éparpillés autour de la ligne 0. Il ne semble donc pas y avoir de soucis au niveau du postulat de linéarité.

Normalité :

Dans le graphique en haut à droite de la **figure 8**, les résidus studentisés suivent relativement bien la droite normale excepté qu'ils ont une queue lourde entre les quantiles théoriques 2 et 4.

Homoscédasticité :

Dans le graphique en bas à gauche de la **figure 8**, les résidus studentisés semblent constants peu importe la valeur de ŷ.

Détection de valeurs aberrantes :

Dans le graphique en bas à droite de la **figure 8**, il ne semble pas y avoir de problème de valeurs aberrantes, car aucun des points se trouvent en dehors du seuil critique de la distance de Cook.