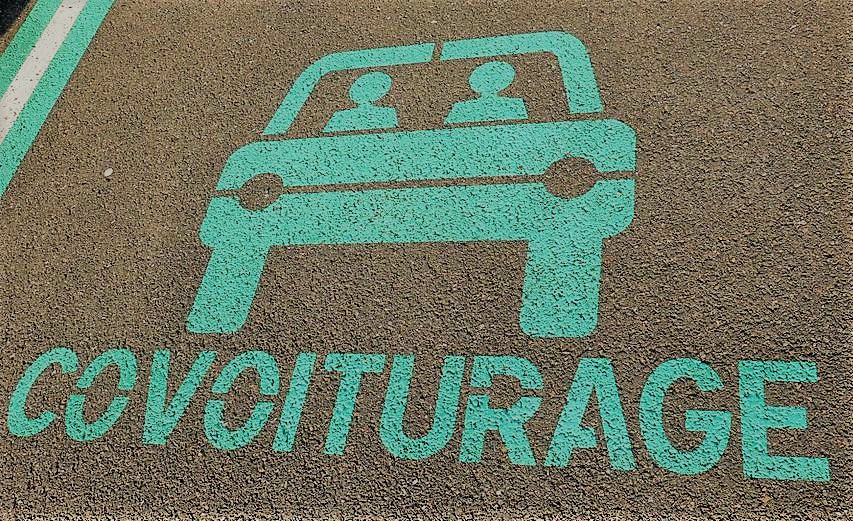
Covoiturage - L’offre et la demande

cb

04/12/2020



## Outils

knitr::opts\_chunk$set(echo = TRUE)  
library(readr)  
library(tidyverse)  
library(reshape2)  
library(gridExtra)  
library(lme4)  
library(jtools)  
library(interactions)

## Données et recodage

C’est un vieux, déja, jeu de données scrappées sur un site de covoiturage. Environ 13000 offres proposées sur différents types de trajets et de destination à un moment donné (au cours de 2016). Je voulais à l’époque tester l’effet des signaux de qualités sur les comportements dans l’environnement collaboratif. S’agit-il d’un marché ou d’un réseau d’entraide ?

Si elle ne sont plus d’actualité, elle permettent cependant une étude intéressante des facteurs qui encourage la demande mais aussi d’analyser des facteurs d’offre et noatmment les déterminant du prix. A cette période les condcuteurs avaient le loisir de fixer leur prix même si une recommandation était donnée par la plateforme. On sera en particulier attentif au rôle du système de réputation dont Blablacar a fait le coeur de son modèles (cf DREAMS).

df <- read\_delim("Blablav2.csv", ";", escape\_double = FALSE, trim\_ws = TRUE)

## Warning: Missing column names filled in: 'X19' [19], 'X20' [20], 'X21' [21],  
## 'X22' [22]

## Parsed with column specification:  
## cols(  
## .default = col\_character(),  
## Num = col\_double(),  
## Distance = col\_double(),  
## `Temps mn` = col\_double(),  
## Nombre = col\_double(),  
## `Places Restantes` = col\_double(),  
## Age = col\_double(),  
## Note = col\_double(),  
## Prix = col\_double(),  
## X21 = col\_logical(),  
## X22 = col\_logical()  
## )

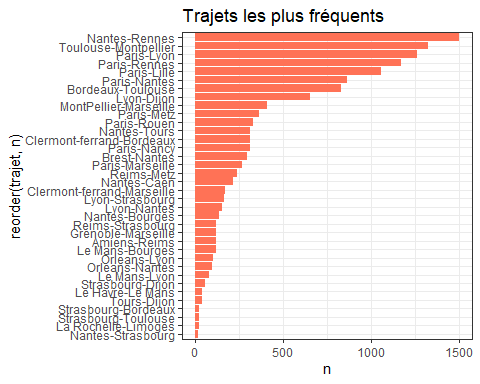
## See spec(...) for full column specifications.

## Warning: 13 parsing failures.  
## row col expected actual file  
## 1644 X21 1/0/T/F/TRUE/FALSE Europe, Aix-en-Prov : voir avec le conducteur 'Blablav2.csv'  
## 1644 X22 1/0/T/F/TRUE/FALSE Depart pile a l'heure 'Blablav2.csv'  
## 2386 Age a double "32 - <strong class=""green"">Helper</strong>" 'Blablav2.csv'  
## 2788 Age a double "36 - <strong class=""green"">Helper</strong>" 'Blablav2.csv'  
## 3561 Age a double "28 - <strong class=""green"">Helper</strong>" 'Blablav2.csv'  
## .... ... .................. .............................................. ..............  
## See problems(...) for more details.

df<- df%>% rename(TauxOccup=`Places Restantes`,   
 Depart=`Ville Depart Capture`,  
 Arrivee=`Ville Arrivee Capture`,  
 Flex=`Flexibilite Horaire`  
  
 )  
df<-df %>% filter( Distance<1000)  
  
#on recode les places restantes par une approximation du taux de réservation  
df$Occup<- NA  
df$Occup[df$TauxOccup==0]<-1  
df$Occup[df$TauxOccup==1]<-.75  
df$Occup[df$TauxOccup==2]<-.50  
df$Occup[df$TauxOccup==3]<-.25  
df$Occup[df$TauxOccup==4]<-0  
df$Occup[df$TauxOccup>=4]<-0  
  
#l'expérience peut se capter par le nombre de voyages  
df$Nombre[is.na(df$Nombre)]<-0

Voici la liste des requêtes et le nombre d’offres obtenues pour chacune d’elle. On observe la domination des trajet inter-régionaux avec NantesRennes , toulouze Montpellier et Bordeaux Toulouse.

trajet<-table(df$Depart, df$Arrivee)  
  
df$trajet<-paste0(df$Depart,"-",df$Arrivee)  
foo<-df %>% mutate(n=1)%>%group\_by(trajet)%>%summarise(n=sum(n))%>%filter(n>0)  
  
ggplot(foo,aes(x=reorder(trajet,n), y=n))+geom\_bar(stat="identity", fill="coral1")+coord\_flip()+theme\_bw()+labs(title="Trajets les plus fréquents")

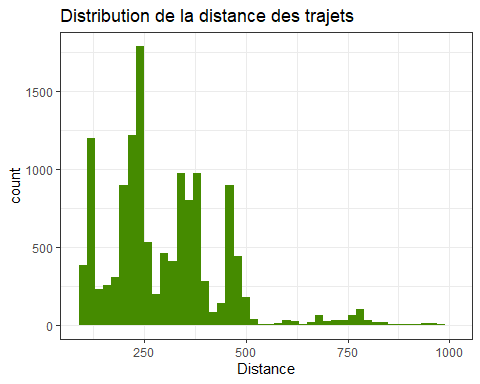


## Un tour d’horizon des déterminants ptentiel du prix et des taux d’occupation.

### La distance

Les trajets ont été échantillonnés pour représenter différents niveaux d’échelle de distance, les cours trajets, les trajets de 100 à 200 km, ceux à 300-400 et une minorité de distance plus longue, ce qui explique la concentration à certains niveaux. LA distribution ne représente pas la distribution de la demande des trajets, ni même l’offre. L’échantillon est en fait un pool d’offres pour un trajet donné.

ggplot(df, aes(x=Distance))+geom\_histogram(fill="Chartreuse4",binwidth = 20)+theme\_bw() +labs(title = "Distribution de la distance des trajets")



### La voiture

à traiter et d’abord recoder. Rôle de la marque et du standing du véhicule

# ggplot(df, aes(x=Marque))+geom\_bar(fill="Chartreuse4")+theme\_bw() +coord\_flip()

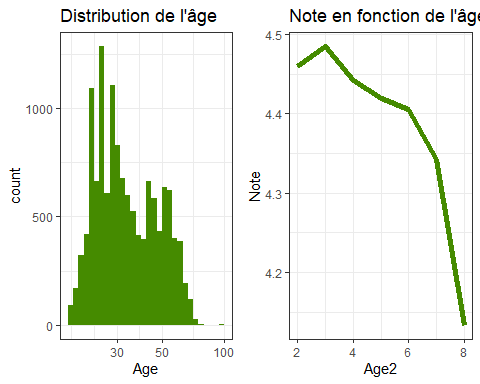
### L’âge du capitaine

On distingue deux populations en terme d’âge, une en dessous de la trentaine, l’autre de 40 à 50 ans. On Jette un coup d’oeil ensuite sur la relation entre l’âge et la note qui culmine à 30 ans et baisse avec les décades (F=59.3). Est-ce l’effet d’une inadéquation des âges? La demande est-elle plus jeune que l’offre? Celà crée-t-il un biais systématique d’évaluation?

g2<-ggplot(df, aes(x=Age))+geom\_histogram(fill="Chartreuse4")+theme\_bw() +scale\_x\_log10() +labs(title="Distribution de l'âge")  
df$Age2<-(round(df$Age/10,0))  
foo<-df %>% group\_by(Age2)%>%summarise(Note=mean(Note, na.rm=TRUE)) %>% filter(Age2<9)   
g3<-ggplot(foo, aes(x=Age2, y=Note, group=1))+geom\_line(stat="identity",color="Chartreuse4",size =2)+theme\_bw()+labs(title="Note en fonction de l'âge")  
  
grid.arrange(g2,g3,ncol=2)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 13 rows containing non-finite values (stat\_bin).



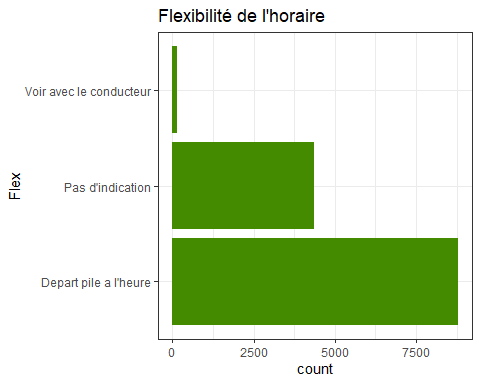
fit<-lm(Note~Age2, df)  
anova(fit)

## Analysis of Variance Table  
##   
## Response: Note  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Age2 1 6.47 6.4686 59.325 1.444e-14 \*\*\*  
## Residuals 11732 1279.21 0.1090   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### Le trajet

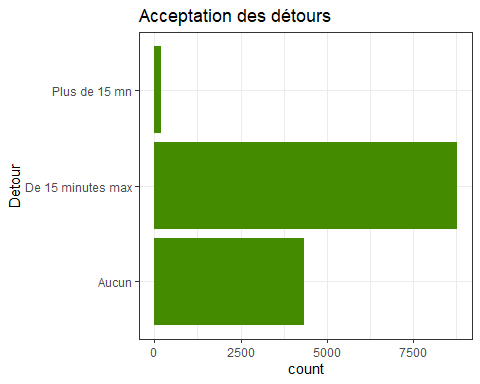
Le conducteur peut être flexible dans l’horaire de départ ou pas. Un petit recodage est cependant néçessaire.

df$Flex[df$Flex!="Depart pile a l'heure" & df$Flex!="Pas d'indication"]<-"Voir avec le conducteur"  
df$Flex[is.na(df$Flex)]<-"Pas d'indication"  
  
ggplot(df, aes(x=Flex))+geom\_bar(fill="Chartreuse4")+theme\_bw() +coord\_flip()+labs(title="Flexibilité de l'horaire")



Le trajet peut être strict ou comporter des détours. Là aussi besoin d’un peu de recodage.

df$Detour[df$Detour=="De 30 minutes max"]<-"Plus de 15 mn"  
df$Detour[df$Detour=="Autant que possible"]<-"Plus de 15 mn"  
ggplot(df, aes(x=Detour))+geom\_bar(fill="Chartreuse4")+theme\_bw() +coord\_flip()+labs(title="Acceptation des détours")

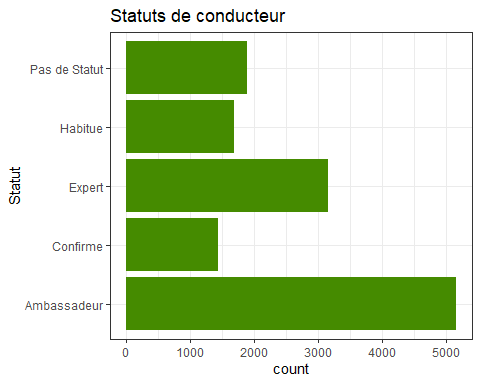


### le conducteur : les notes et l’expérience et statut

Le conducteur se manifeste au travers de 3 critères : son statut attribué par blablacar, la note moyenne reçue des passagers et son expérience traduite par le nombre de voyage qu’il a fait.

Examinons le statut et notamment en comparant l’expérience et l’évaluation.

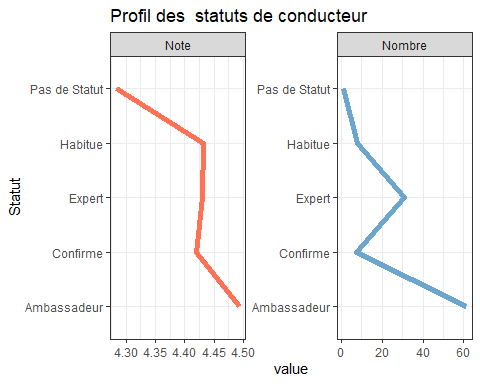
df$Statut<-as.factor(df$Statut)  
ggplot(df, aes(x=Statut))+  
 geom\_bar(fill="Chartreuse4")+theme\_bw() +  
 coord\_flip()+  
 labs(title = "Statuts de conducteur")



foo<-df %>% group\_by(Statut)%>%summarise(Note=mean(Note,na.rm=TRUE),Nombre=mean(Nombre,na.rm=TRUE))  
foo<-melt(foo)

## Using Statut as id variables

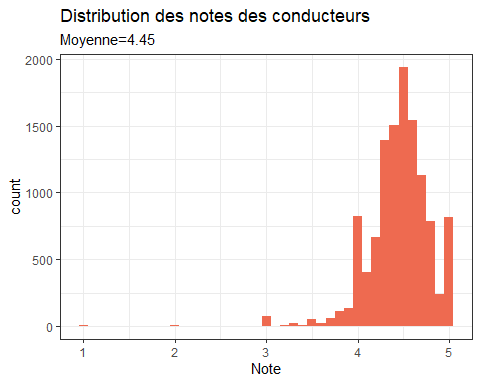
ggplot(foo, aes(x=Statut, y=value,group=variable))+geom\_line(aes(color=variable), size=2)+  
 theme\_bw() +  
 coord\_flip()+facet\_wrap(vars(variable),scales ="free")+ labs(title = "Profil des statuts de conducteur")+scale\_color\_manual(values=c("coral1", "skyblue3"))+ guides(color = FALSE, size = FALSE)



les notes sont en moyenne de 4.45 et fortement déviées à droite, réduisant la capacité de discrimination.

mean<-round(mean(df$Note,na.rm=TRUE),2)  
ggplot(df, aes(x=Note))+geom\_histogram(fill="coral2",binwidth = .1)+  
 theme\_bw() +  
 labs(title = "Distribution des notes des conducteurs", subtitle =paste0("Moyenne=", mean))

## Warning: Removed 1539 rows containing non-finite values (stat\_bin).

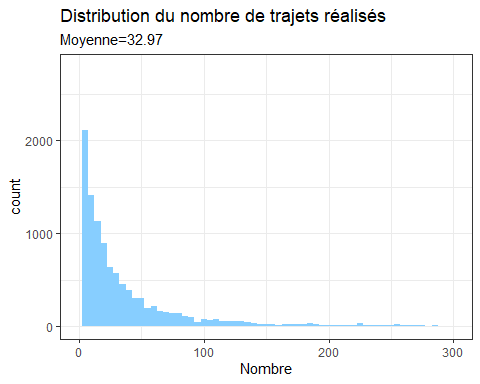


L’expérience se distribue de manière très inégale, une minorité ayant réalisé plus de 50 voyages, et donc on intégré l’usage de manière quasi hebdmadaire voir quotidienne, une très grande majorité en ayant fait moins d’une vingtaine.

mean<-round(mean(df$Nombre,na.rm=TRUE),2)  
  
ggplot(df, aes(x=Nombre))+geom\_histogram(fill="skyblue1", binwidth = 5)+theme\_bw()+labs(title = "Distribution du nombre de trajets réalisés", subtitle =paste0("Moyenne=", mean))+xlim(0,300)

## Warning: Removed 106 rows containing non-finite values (stat\_bin).

## Warning: Removed 2 rows containing missing values (geom\_bar).

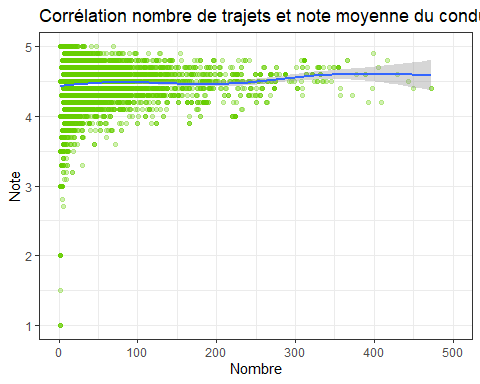
 Note et exprérience sont-ils corrélés? Rien de très probant même s’il semble que la note s’accroisse légèrement avec le nombre de voyages réalisés.

ggplot(df, aes(x=Nombre, y=Note))+geom\_point(color="Chartreuse3",alpha=0.3)+theme\_bw() +geom\_smooth()+xlim(0,500)+labs(title="Corrélation nombre de trajets et note moyenne du conducteur")

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

## Warning: Removed 1561 rows containing non-finite values (stat\_smooth).

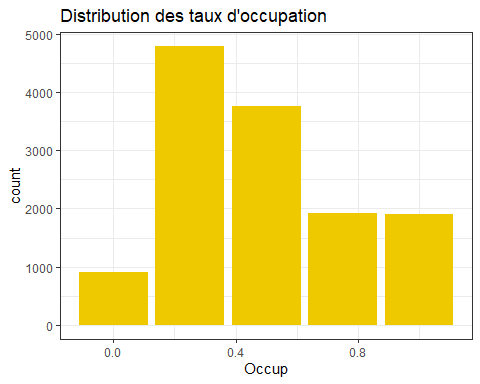
## Warning: Removed 1561 rows containing missing values (geom\_point).



## Analyser la demande

On utilise la variable nombre de place restante qu’on traduit par un indicateur codé de 0 (tout est libre) à 1 ( la voiture est pleine). En voici la distribution.

ggplot(df, aes(x=Occup))+geom\_bar(fill="gold2")+theme\_bw() +labs(title="Distribution des taux d'occupation")



### Un modèle OLS

On commence par un modèle simple et linéaire où l’on cherche à expliquer, prédire, le taux d’occupation du véhicule en fonction des variables dont nous disposons. La flexibilité de l’horaire et la possibilité de détour n’affecte pas vraiment le taux de réservation. Si les autres variables ont des relations significatives ( avec des valeur t très élevées), on notera que la variance expliquée est faible.

Le remplissage des voiture est une affaire de loto. Ce qui se comprend, car la probabilité qu’une offre et une demande coincide est relativement faible si on retient une plage horaire étroite pour le passager qui l’élargira pour accroître son choix au prix d’un effort supplémentaire de recherche.

fit0<-lm(Occup~Age+Distance+Prix+Flex+Detour+Autoroute+Statut+log(Nombre+1)+Note, data=df)  
summ(fit0)

Observations

11706 (1579 missing obs. deleted)

Dependent variable

Occup

Type

OLS linear regression

F(14,11691)

19.88

R²

0.02

Adj. R²

0.02

Est.

S.E.

t val.

p

(Intercept)

0.20

0.04

4.60

0.00

Age

-0.00

0.00

-4.26

0.00

Distance

0.00

0.00

8.96

0.00

Prix

-0.01

0.00

-8.39

0.00

FlexPas d’indication

0.01

0.01

1.81

0.07

FlexVoir avec le conducteur

0.01

0.03

0.32

0.75

DetourDe 15 minutes max

-0.00

0.01

-0.47

0.64

DetourPlus de 15 mn

-0.07

0.03

-2.39

0.02

AutorouteOui

0.11

0.02

6.97

0.00

StatutConfirme

-0.03

0.01

-2.81

0.00

StatutExpert

-0.02

0.01

-2.72

0.01

StatutHabitue

-0.05

0.01

-4.47

0.00

StatutPas de Statut

-0.07

0.02

-3.61

0.00

log(Nombre + 1)

-0.02

0.00

-6.59

0.00

Note

0.06

0.01

7.32

0.00

Standard errors: OLS

Le modèle précédent est basique, nous allons le reformuler. Nous écartons les variables peu significatives (en termes de t<3) pour se concentrer sur la distance du trajet, le choix de l’autoroute qui garantit la rapidité du voyage, l’âge du répondant et surtout la note et le nombre de trajets qui signalent l’expérience du conducteur. Ce sont des éléments clés du système de réputation.

Les deux modèles sont représentés dans un diagramme de paramètres qui permet de mieux illustrer les résultats qu’un tableau classique de paramètres de régression. On emploie jtools pour les représentations graphiques/

fit1<-lm(Occup~Age+Statut+Distance+Prix+Autoroute+log(Nombre+1)+Note:log(Nombre+1), data=df)  
summ(fit1)

Observations

11706 (1579 missing obs. deleted)

Dependent variable

Occup

Type

OLS linear regression

F(10,11695)

29.18

R²

0.02

Adj. R²

0.02

Est.

S.E.

t val.

p

(Intercept)

0.46

0.02

21.34

0.00

Age

-0.00

0.00

-4.04

0.00

StatutConfirme

-0.03

0.01

-2.66

0.01

StatutExpert

-0.02

0.01

-2.37

0.02

StatutHabitue

-0.05

0.01

-4.33

0.00

StatutPas de Statut

-0.07

0.02

-3.77

0.00

Distance

0.00

0.00

8.97

0.00

Prix

-0.01

0.00

-8.46

0.00

AutorouteOui

0.11

0.02

7.11

0.00

log(Nombre + 1)

-0.16

0.02

-9.95

0.00

log(Nombre + 1):Note

0.03

0.00

8.76

0.00

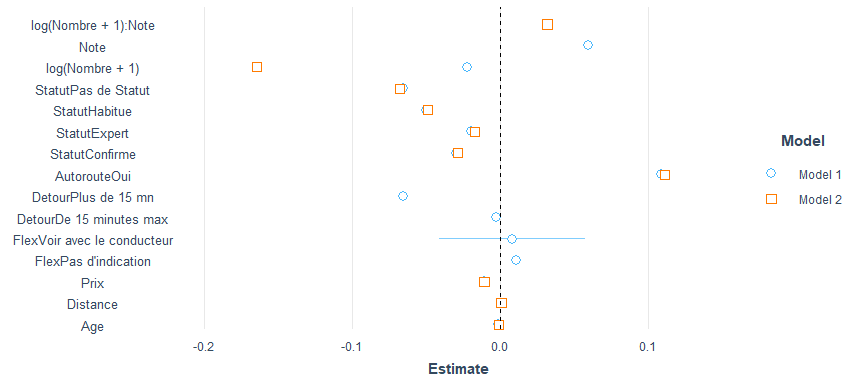
Standard errors: OLS

plot\_summs(fit0,fit1, plot.distributions = TRUE,scale = FALSE)

## Loading required namespace: broom.mixed

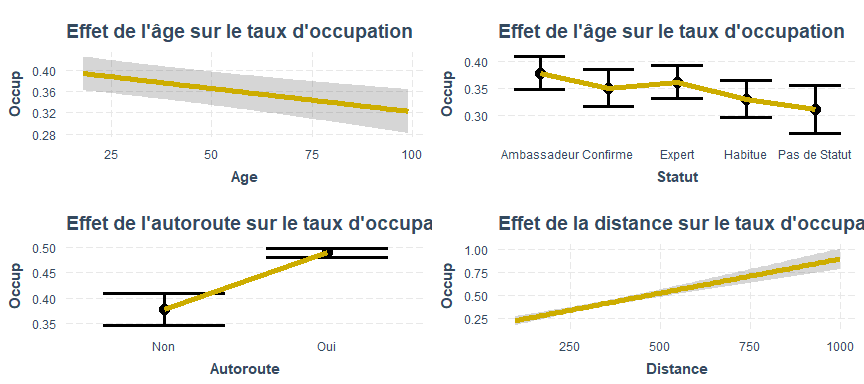
## Registered S3 method overwritten by 'broom.mixed':  
## method from   
## tidy.gamlss broom

## If some of the distribution curves are too short to see, consider rescaling  
## your model coefficients or using the rescale.distributions = TRUE argument.

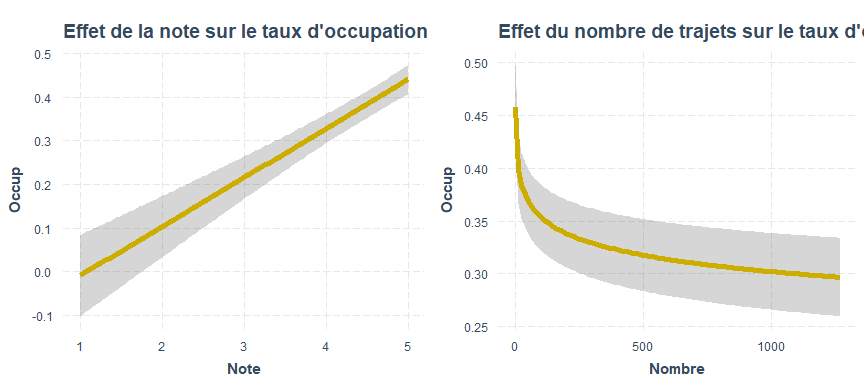


Le modèle tient même s’il explique très peu de variance. Les tests sont cependant significatifs et même si l’effet est faible il est notable. Voici des diagrammes d’effets qui représentent la valeur de la variable dépendante pour une plage de variation des variables indépendantes prises une à une, en fonction du modèle. On voit ainsi plus clairement que le taux d’occupation passe de 20% environ quand la note est de 3 (les notes inférieures sont rares!) à 45% quand la moyenne du conducteur frôle les 5.

g10<-effect\_plot(fit1, pred=Age,interval = TRUE ,data=df)+geom\_line(color="gold3",size=2)+labs(title="Effet de l'âge sur le taux d'occupation")  
g11<-effect\_plot(fit1, pred=Statut,interval = TRUE ,data=df)+geom\_line(color="gold3",size=2)+labs(title="Effet de l'âge sur le taux d'occupation")  
g12<-effect\_plot(fit1, pred=Autoroute,interval = TRUE ,data=df)+geom\_line(color="gold3",size=2)+labs(title="Effet de l'autoroute sur le taux d'occupation")  
g13<-effect\_plot(fit1, pred=Distance,interval = TRUE ,data=df)+geom\_line(color="gold3",size=2)+labs(title="Effet de la distance sur le taux d'occupation")  
grid.arrange(g10,g11,g12,g13,ncol=2)



g14<-effect\_plot(fit1, pred=Note,interval = TRUE ,data=df)+geom\_line(color="gold3",size=2)+labs(title="Effet de la note sur le taux d'occupation")  
g15<-effect\_plot(fit1, pred=Nombre,interval = TRUE ,data=df)+geom\_line(color="gold3",size=2)+labs(title="Effet du nombre de trajets sur le taux d'occupation")  
grid.arrange(g14,g15,ncol=2)

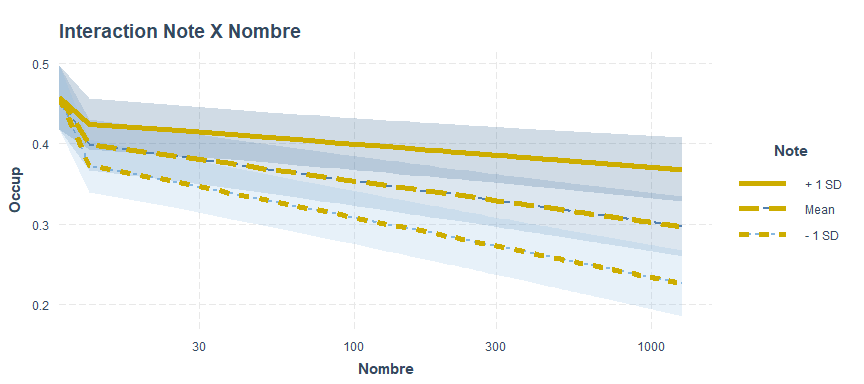


La particularité du modèle est le terme d’interaction. L’expérience du conducteur traduite par le nombre de trajets qu’il a effectué est gage de sécurité, de confiance, pourvu que ses notes soient bonnes. On s’attend à ce que si elle sont moins bonnes, le nombre de voyages réalisés en amplifiera l’impact négatif. L’expérience signale aussi la crédibilité de la note, et on peut raisonnablement penser que plus ce nombre est grand et plus le signal, positif ou négatif est crédible. Ici la note module l’effet de crédibilité.

L’analyse de l’interaction est claire : quand les contenus sont négatifs , l’expérience du conducteur aggrave la réticence et le taux d’occupation se réduit. Quand les notes vont meilleures que la moyenne.L’amplification par l’expérience de l’effet des notes sur la réservation est moindre. L’interaction est significative (t=8.76). Reste à comprendre pourquoi la tendance générale reste à une relation négative entre le taux de remplissage et le nombre de trajet effectué . Est-ce le résultat d’une participation fréquente à la plateforme, à un comportement particulier des conducteurs très actifs (opportunisme) ? Une explication complémentaire peut venir de ceux qui pratique occasionnellement le co-voiturage, leur activité étant plus rare , elle est peut être plus plannifiée, et les offres sont présentes depuis plus longtemps que celles des utilisateurs fréquents qui publiraient leurs annonces de la veille au lendemain. Ceci mérite une analyse plus approfondie que nous ne menerons pas ici.

library(interactions)  
interact\_plot(fit1,pred =Nombre , modx =Note ,interval = TRUE ,data=df) +scale\_x\_log10()+geom\_line(color="gold3",size=2)+labs(title="Interaction Note X Nombre")

## Warning: Transformation introduced infinite values in continuous x-axis  
  
## Warning: Transformation introduced infinite values in continuous x-axis  
  
## Warning: Transformation introduced infinite values in continuous x-axis



### Un modèle plus raffiné

On apporte deux ajustements :

* La spécification du modèle est logistique, car le taux d’occupation va de 0 à 100% même si nous n’avons que 4 niveaux. ça fait un modèle plus réaliste qui ajuste la valeur prédite entre 0 et 1.
* Pour mieux prendre en compte l’hterogénéité du set de données (composée par un échantillon de requêtes) on introduit les trajets comme composante aléatoire.

l’estimation est réalisée avec lme4

Les effets semblent être ici plus pertinent et sensibles mais surtout l’effet d’interaction semble renforcé. L’effet de la note est là clairement amplifié pour les notes négatives mais aussi pour les notes positives.

fit2<-glmer(Occup~Age+Statut+Distance+Prix+Autoroute+log(Nombre+1)+Note:log(Nombre+1)+(1|trajet),family = binomial, data=df)

## Warning in eval(family$initialize, rho): non-integer #successes in a binomial  
## glm!

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :  
## Model failed to converge with max|grad| = 9.27091 (tol = 0.002, component 1)

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, : Model is nearly unidentifiable: very large eigenvalue  
## - Rescale variables?;Model is nearly unidentifiable: large eigenvalue ratio  
## - Rescale variables?

summ(fit2)

Observations

11706

Dependent variable

Occup

Type

Mixed effects generalized linear model

Family

binomial

Link

logit

AIC

13913.06

BIC

14001.47

Pseudo-R² (fixed effects)

0.02

Pseudo-R² (total)

0.02

Fixed Effects

Est.

S.E.

z val.

p

(Intercept)

-1.10

0.18

-6.20

0.00

Age

-0.01

0.00

-3.00

0.00

StatutConfirme

-0.18

0.08

-2.19

0.03

StatutExpert

-0.17

0.06

-3.09

0.00

StatutHabitue

-0.39

0.09

-4.42

0.00

StatutPas de Statut

-0.16

0.14

-1.14

0.25

Distance

0.01

0.00

8.17

0.00

Prix

-0.08

0.01

-7.89

0.00

AutorouteOui

0.62

0.13

4.65

0.00

log(Nombre + 1)

-0.77

0.13

-5.85

0.00

log(Nombre + 1):Note

0.15

0.03

5.39

0.00

Random Effects

Group

Parameter

Std. Dev.

trajet

(Intercept)

0.00

Grouping Variables

Group

# groups

ICC

trajet

36

0.00

g16a<-effect\_plot(fit2, pred=Age,interval = TRUE ,data=df)+geom\_line(color="gold3",size=2)+labs(title="Effet de l'âge du conducteur",y="taux d'occupation" )

## Confidence intervals for merMod models is an experimental feature. The  
## intervals reflect only the variance of the fixed effects, not the random  
## effects.

g16b<-effect\_plot(fit2, pred=Statut,interval = TRUE ,data=df)+geom\_line(color="gold3",size=2)+labs(title="Effet du statut du conducteur",y="taux d'occupation")

## Confidence intervals for merMod models is an experimental feature. The  
## intervals reflect only the variance of the fixed effects, not the random  
## effects.

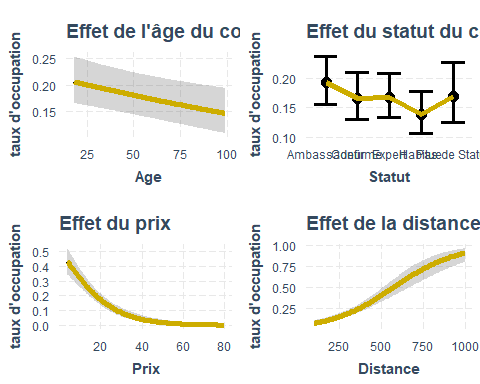
g16<-effect\_plot(fit2, pred=Prix,interval = TRUE ,data=df)+geom\_line(color="gold3",size=2)+labs(title="Effet du prix",y="taux d'occupation")

## Confidence intervals for merMod models is an experimental feature. The  
## intervals reflect only the variance of the fixed effects, not the random  
## effects.

g17<-effect\_plot(fit2, pred=Distance,interval = TRUE ,data=df)+geom\_line(color="gold3",size=2)+labs(title="Effet de la distance du trajet",y="taux d'occupation")

## Confidence intervals for merMod models is an experimental feature. The  
## intervals reflect only the variance of the fixed effects, not the random  
## effects.

grid.arrange(g16a,g16b,g16,g17,ncol=2)



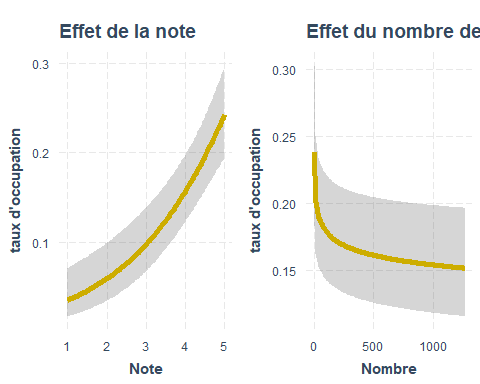
g18<-effect\_plot(fit2, pred=Note,interval = TRUE ,data=df)+geom\_line(color="gold3",size=2)+labs(title="Effet de la note",y="taux d'occupation")

## Confidence intervals for merMod models is an experimental feature. The  
## intervals reflect only the variance of the fixed effects, not the random  
## effects.

g19<-effect\_plot(fit2, pred=Nombre,interval = TRUE ,data=df)+geom\_line(color="gold3", size=2)+labs(title="Effet du nombre de trajets",y="taux d'occupation")

## Confidence intervals for merMod models is an experimental feature. The  
## intervals reflect only the variance of the fixed effects, not the random  
## effects.

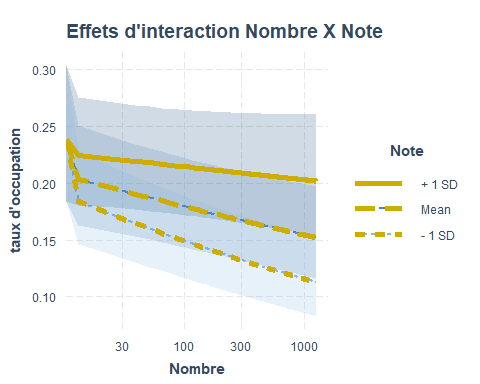
grid.arrange(g18,g19,ncol=2)



g20<-interact\_plot(fit2,pred =Nombre , modx =Note ,interval = TRUE ,data=df) +scale\_x\_log10()+geom\_line(color="gold3", size=2)+labs(title="Effets d'interaction Nombre X Note",y="taux d'occupation")  
g20

## Warning: Transformation introduced infinite values in continuous x-axis

## Warning: Transformation introduced infinite values in continuous x-axis  
  
## Warning: Transformation introduced infinite values in continuous x-axis



## L’offre et donc le prix

Le prix est un facteur important, du moins parmi ceux donc le conducteur dispose pour remplir son véhicule et optimiser son revenu.

On s’attend de manière générale à ce que si le conducteur est correctement signalé : il a de bonne note et une grande expérience, à ce qu’il tire avantage de cette qualification en accroissant le prix qu’il offre par rapport à d’autre conducteurs moins aguérris au système et moins bien évalué.

L’hypothèse symétrique doit aussi être considérée : celle qui considère que les conducteurs expérimenté offre des prix plus faibles pour bénéficier d’un avantage de réputation. Si les notes ont un effet positif sur le remplissage du véhicule, il est peut-être avantageux de baisserles prix.

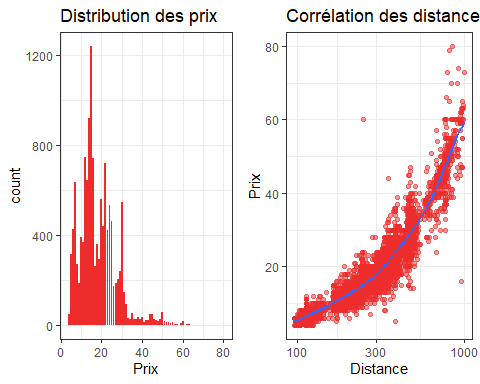
Examinons les faits.

## Analyse des prix

Au premier examen la distribution des prix semble être multimodale, en fait elle est étroitement associée à la distribution des distance. Le prix est une multiplication de la distance par un tarif kilométrique, même si la relation ne semble pas tout à faire linéaire.

g20<-ggplot(df, aes(x=Prix))+geom\_bar(fill="firebrick2")+theme\_bw() +labs(title = "Distribution des prix")  
g21<-ggplot(df, aes(x=Distance, y=Prix))+geom\_point(color="firebrick2", alpha =0.5)+theme\_bw() +geom\_smooth(method="gam")+scale\_x\_log10()+labs(title = "Corrélation des distances et des prix")  
library(gridExtra)  
grid.arrange(g20,g21, ncol=2)

## `geom\_smooth()` using formula 'y ~ s(x, bs = "cs")'



C’est pourquoi on calcule un tarif au km, qui lui n’est plus corrélé ou presque à la distance parcourue.

df$prix\_km<-df$Prix/df$Distance  
  
g22<-ggplot(df, aes(x=prix\_km))+geom\_histogram(fill="firebrick3", binwidth = 0.005)+theme\_bw() +labs(title = "Distribution des prix")+xlim(0,0.15)  
g23<-ggplot(df, aes(x=Distance, y=prix\_km))+geom\_point(color="firebrick3")+theme\_bw() +geom\_smooth(method="gam")+scale\_x\_log10()+ylim(0,0.15)  
grid.arrange(g22,g23, ncol=2)

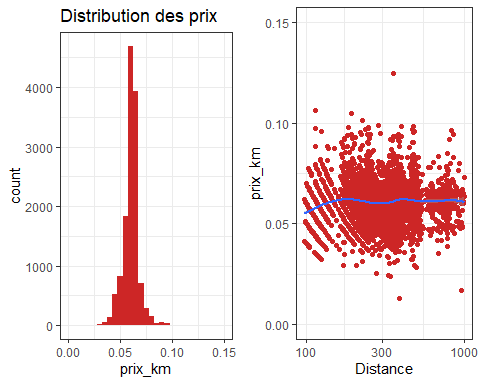
## Warning: Removed 1 rows containing non-finite values (stat\_bin).

## Warning: Removed 2 rows containing missing values (geom\_bar).

## `geom\_smooth()` using formula 'y ~ s(x, bs = "cs")'

## Warning: Removed 1 rows containing non-finite values (stat\_smooth).

## Warning: Removed 1 rows containing missing values (geom\_point).



## Analyse du prix

On commence par un modèle simple et linéaire où l’on cherche à expliquer, prédire, le taux d’occupation du véhicule en fonction de trois variables:

fit3<-lm(prix\_km~Age+Distance+Autoroute+log(Nombre+1)+Note, data=df)  
summ(fit3)

Observations

11706 (1579 missing obs. deleted)

Dependent variable

prix\_km

Type

OLS linear regression

F(5,11700)

179.58

R²

0.07

Adj. R²

0.07

Est.

S.E.

t val.

p

(Intercept)

0.06

0.00

56.91

0.00

Age

0.00

0.00

9.18

0.00

Distance

0.00

0.00

11.30

0.00

AutorouteOui

0.01

0.00

22.27

0.00

log(Nombre + 1)

-0.00

0.00

-9.61

0.00

Note

-0.00

0.00

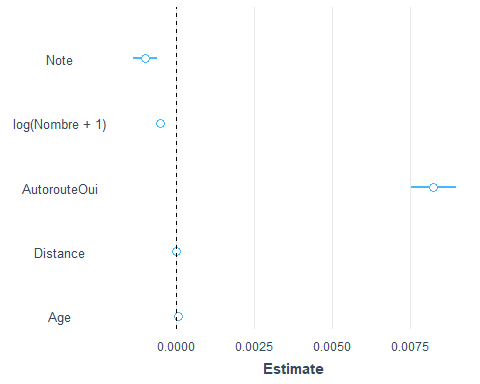
-5.15

0.00

Standard errors: OLS

plot\_summs(fit3, plot.distributions = TRUE,scale = FALSE)

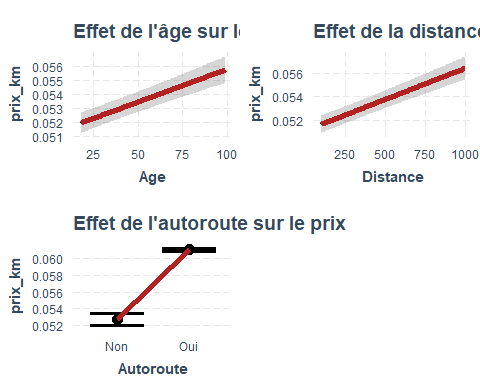
## If some of the distribution curves are too short to see, consider rescaling  
## your model coefficients or using the rescale.distributions = TRUE argument.



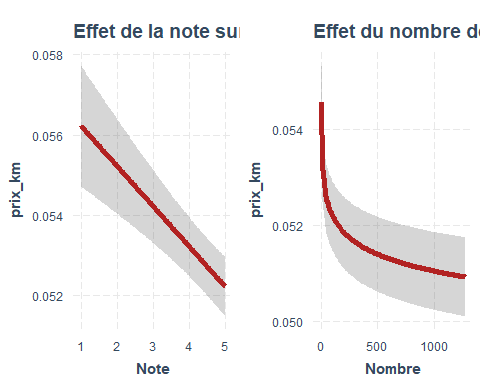
g24<-effect\_plot(fit3, pred=Age,interval = TRUE ,data=df)+geom\_line(color="firebrick",size=2)+labs(title="Effet de l'âge sur le prix")  
g25<-effect\_plot(fit3, pred=Distance,interval = TRUE ,data=df)+geom\_line(color="firebrick",size=2)+labs(title="Effet de la distance sur le prix")  
g27<-effect\_plot(fit3, pred=Autoroute,interval = TRUE ,data=df)+geom\_line(color="firebrick",size=2)+labs(title="Effet de l'autoroute sur le prix")  
  
g26<-effect\_plot(fit3, pred=Note,interval = TRUE ,data=df)+geom\_line(color="firebrick",size=2)+labs(title="Effet de la note sur le prix")  
g28<-effect\_plot(fit3, pred=Nombre,interval = TRUE ,data=df)+geom\_line(color="firebrick",size=2)+labs(title="Effet du nombre de trajets")  
g29<-interact\_plot(fit3,pred =Nombre , modx =Note ,interval = TRUE ,data=df) +scale\_x\_log10()+geom\_line(color="firebrick",size=2)+labs(title="Interaction nombre X Note")

## Warning: Nombre and Note are not included in an interaction with one another in the  
## model.

grid.arrange(g24,g25,g27, ncol=2)



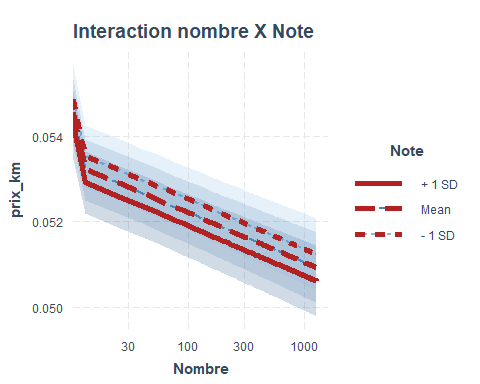
grid.arrange(g26,g28, ncol=2)



g29

## Warning: Transformation introduced infinite values in continuous x-axis

## Warning: Transformation introduced infinite values in continuous x-axis  
  
## Warning: Transformation introduced infinite values in continuous x-axis



On ajoute l’effet des départs et des arrivées en composantes d’erreur pour capturer l’hétérogénéité.

Ici par d’interaction :

* plus l’expérience est grande et plus le prix au km est faible. Il est un peu plus élevé si la note est meilleure.
* meilleure est la note et plus élevé est le prix au km

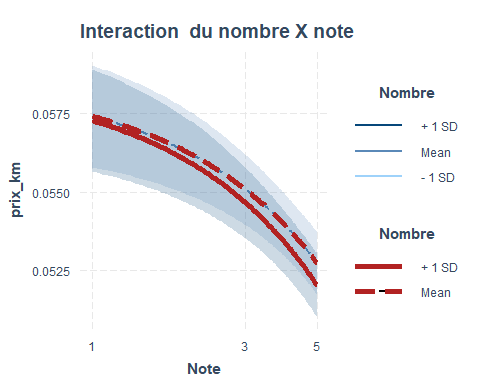
Les conducteurs valorisent leurs réputation, mais les plus expérimentés ont des stratégies de prix bas. Est-ce un calcul pour accelérer les réservations de place ? ou est-ce le reflet de leur engagement : en baissant les prix ils signalent qu’il ne veulent pas abuser de leur position. C’est difficile de trancher, mais le résultat est là.

fit4<-lmer(prix\_km~Distance+Autoroute+Age+Note+log(Nombre+1):Note+(1|trajet), data=df)  
summary(fit4)

## Linear mixed model fit by REML ['lmerMod']  
## Formula: prix\_km ~ Distance + Autoroute + Age + Note + log(Nombre + 1):Note +   
## (1 | trajet)  
## Data: df  
##   
## REML criterion at convergence: -83637.9  
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -6.1783 -0.4504 0.0499 0.4561 9.1001   
##   
## Random effects:  
## Groups Name Variance Std.Dev.  
## trajet (Intercept) 3.046e-06 0.001745  
## Residual 4.542e-05 0.006739  
## Number of obs: 11706, groups: trajet, 36  
##   
## Fixed effects:  
## Estimate Std. Error t value  
## (Intercept) 5.787e-02 1.133e-03 51.093  
## Distance -2.858e-06 1.471e-06 -1.943  
## AutorouteOui 8.157e-03 3.978e-04 20.505  
## Age 4.381e-05 4.996e-06 8.769  
## Note -6.793e-04 1.926e-04 -3.527  
## Note:log(Nombre + 1) -1.370e-04 1.157e-05 -11.839  
##   
## Correlation of Fixed Effects:  
## (Intr) Distnc AutrtO Age Note   
## Distance -0.463   
## AutorouteOu -0.319 -0.042   
## Age -0.201 -0.019 -0.006   
## Note -0.746 0.000 -0.002 0.085   
## Nt:lg(Nm+1) -0.013 0.054 0.035 -0.072 -0.185

interact\_plot(fit4,pred =Note , modx =Nombre ,interval = TRUE ,data=df) +scale\_x\_log10()+geom\_line(color="firebrick",size=2)+labs(title="Interaction du nombre X note")

## Warning: -27.5026588149518 is outside the observed range of Nombre



## Conclusions

les résultats obtenus montrent d’abord que les échanges dans les plateformes C2C répondent à une logique de marché dans laquelle les stratégies de signaling jouent un rôle certain même s’il est en fait assez marginal au regard des résultats empiriques. Elles expliquent une faible variance des critères retenus : le taux d’occupation du véhicule au moment de la recherche, et le prix proposé par les conducteurs.

Du côté de la demande on onservera que ce sont les caractérisque d’offre qui dominent. Le prix est déterminant autant que la réputation,

Du côté de l’offre on observe cependant une chose étonnante : l’expérience et la qualité de l’évaluation ne se traduisent pas par des comportements de valorisation de la qualité mais au contraire par des stratégie de prix bas.

L’enigme est de savoir pourquoi : s’agit-il simplement d’une stratégie de sécurisation , ceux qui ont l’expérience ont appris qu’en fixant un prix un peu plus bas on s’assure de remplir le véhicule, ou d’une manière de signaler qu’on est un pro du collaboratif et de la générosité et que le prix plus bas signale cette qualité ?

D’autres analyse sont nécessaires sans doute avec une approche plus segmentée à l’instar des deux derniers graphes qui représente les modèles estimés précedemment mais pour chaque groupe de statut.

models <- df %>% group\_by(Statut) %>% summarise(mod = list(lm(Occup~Age+Distance+Autoroute+Prix+log(Nombre+1)+Note:log(Nombre+1))))  
coef5<-as.data.frame(models$mod[[1]]$coefficients)%>%add\_rownames()%>%mutate(Statut="Ambassadeur")%>% rename(coef=2)

## Warning: Deprecated, use tibble::rownames\_to\_column() instead.

coef1<-as.data.frame(models$mod[[2]]$coefficients)%>%add\_rownames()%>% mutate(Statut="Confirmé")%>% rename(coef=2)

## Warning: Deprecated, use tibble::rownames\_to\_column() instead.

coef2<-as.data.frame(models$mod[[3]]$coefficients)%>%add\_rownames()%>%mutate(Statut="Expert")%>% rename(coef=2)

## Warning: Deprecated, use tibble::rownames\_to\_column() instead.

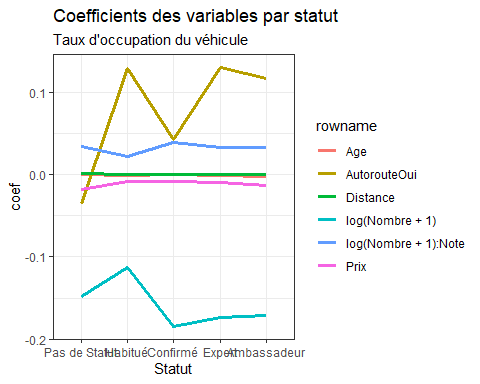
coef3<-as.data.frame(models$mod[[4]]$coefficients)%>%add\_rownames()%>%mutate(Statut="Habitué")%>% rename(coef=2)

## Warning: Deprecated, use tibble::rownames\_to\_column() instead.

coef4<-as.data.frame(models$mod[[5]]$coefficients)%>%add\_rownames()%>%mutate(Statut="Pas de Statut")%>% rename(coef=2)

## Warning: Deprecated, use tibble::rownames\_to\_column() instead.

coef<-rbind(coef1, coef2, coef3, coef4, coef5) %>% filter(rowname!="(Intercept)")  
coef$Statut <- factor(coef$Statut, levels = c("Pas de Statut", "Habitué", "Confirmé", "Expert", "Ambassadeur"))  
  
  
ggplot(coef, aes(x=Statut,y=coef,group=rowname))+geom\_line(aes(color=rowname), size=1.3)+theme\_bw()+labs(title="Coefficients des variables par statut",subtitle = "Taux d'occupation du véhicule")



models <- df %>% group\_by(Statut) %>% summarise(mod = list(lm(prix\_km~Distance+Autoroute+Age+Note+log(Nombre+1):Note)))  
coef5<-as.data.frame(models$mod[[1]]$coefficients)%>%add\_rownames()%>%mutate(Statut="Ambassadeur")%>% rename(coef=2)

## Warning: Deprecated, use tibble::rownames\_to\_column() instead.

coef1<-as.data.frame(models$mod[[2]]$coefficients)%>%add\_rownames()%>% mutate(Statut="Confirmé")%>% rename(coef=2)

## Warning: Deprecated, use tibble::rownames\_to\_column() instead.

coef2<-as.data.frame(models$mod[[3]]$coefficients)%>%add\_rownames()%>%mutate(Statut="Expert")%>% rename(coef=2)

## Warning: Deprecated, use tibble::rownames\_to\_column() instead.

coef3<-as.data.frame(models$mod[[4]]$coefficients)%>%add\_rownames()%>%mutate(Statut="Habitué")%>% rename(coef=2)

## Warning: Deprecated, use tibble::rownames\_to\_column() instead.

coef4<-as.data.frame(models$mod[[5]]$coefficients)%>%add\_rownames()%>%mutate(Statut="Pas de Statut")%>% rename(coef=2)

## Warning: Deprecated, use tibble::rownames\_to\_column() instead.

coef<-rbind(coef1, coef2, coef3, coef4, coef5) %>% filter(rowname!="(Intercept)")  
coef$Statut <- factor(coef$Statut, levels = c("Pas de Statut", "Habitué", "Confirmé", "Expert", "Ambassadeur"))  
  
  
ggplot(coef, aes(x=Statut,y=coef,group=rowname))+geom\_line(aes(color=rowname), size=1.3)+theme\_bw()+labs(title="Coefficients des variables par statut",subtitle = "Prix au km")

