StatWars: models predictius

Arnau Garcia

2024-10-14

StatWars: els meus primers models predictius

En aquesta activitat, us volem presentar una eina molt poderosa que utilitzem en estadística per entendre millor el món que ens envolta: els **models predictius**. Aquests models ens ajuden a predir què passarà en el futur a partir de dades que ja coneixem. A través de conceptes senzills com les **regressions lineals** i les **regressions lineals generalitzades**, veurem com podem utilitzar les matemàtiques per establir relacions entre diferents variables i fer prediccions amb aquestes. Descobrireu com la combinació d'estadística i matemàtiques ens pot ajudar a prendre decisions basades en dades, d'una manera més informada i precisa.

Aquest document serveix com a exemple de com ajustar dos models predictius fent servir el llenguatge de programació **R**. **R**, https://es.wikipedia.org/wiki/R/_(lenguaje_de_programaci%C3%B3n), és el principal llenguatge de programació pels estadístics. Es tracta d'un software lliure, gratuït que es troba a l'abast de tothom.

Veurem dos models predictius:

- Predicció de preus d'habitatges.
- Predicció d'espècies de pingüins.

Predicció de preus d'inmobles

En aquest primer model volem predir el preu de les cases de la ciutat de Ames, a Iowa. Farem servir una base de dades amb dades reals sobre inmobles en aquesta ciutat dels Estats Units. Podeu trobar informació sobre la base de dades original a https://www.openintro.org/data/index.php?data=ames.Farem servir la següent taula de dades per entrenar el nostre primer model predictiu:

```
cases_train <- read.csv("cases_train.csv")
cases_train$X <- NULL
head(cases_train)
```

```
##     preu area qual.total condicio.total  any carrer  tipus garatge.cotxes
## 1 181000 1604          6              5 1997   Pave   1Fam              2
## 2 124000 1584          5              5 1967   Pave Duplex              3
## 3  72000  819          5              4 1919   Pave   1Fam              0
## 4 137000 1229          6              6 1980   Pave TwnhsE              2
## 5 124500  864          4              7 1970   Pave   1Fam              2
## 6 228500 2169          8              5 2002   Pave   1Fam              2
##   garatge.area aire
## 1          470    Y
## 2          792    Y
## 3            0    N
## 4          462    Y
## 5          463    Y
## 6          647    Y
```

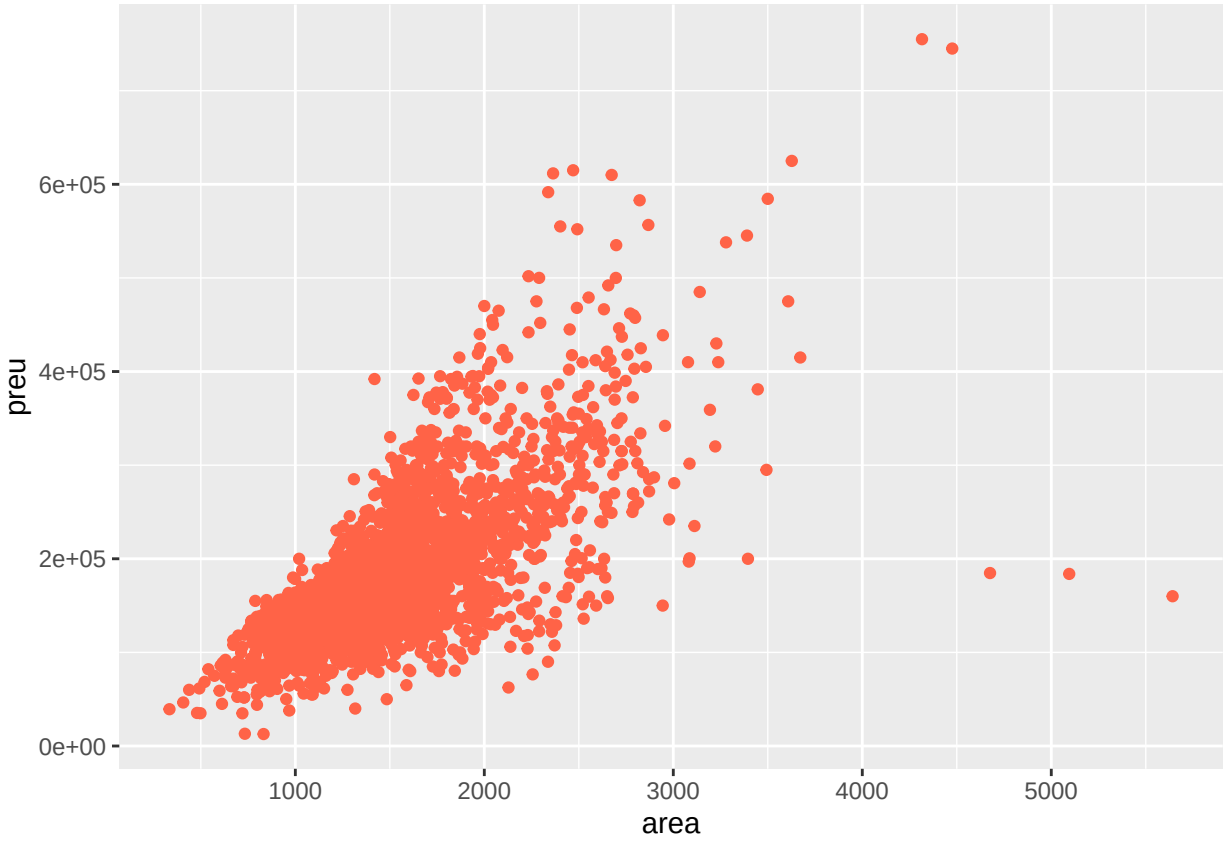
En aquesta base de dades tenim informació sobre 2842 vivendes, per cadascuna tenim 10 variables. Les variables que tenim en aquesta taula són:

- `preu`: preu de l'habitatge.
- `area`: àrea de la vivenda (en peus quadrats).
- `qual.total`: valora de l'1 al 10 el material i el acabat general de l'habitatge.
- `condicio.total`: valora de l'1 al 10 l'estat general de l'habitatge.
- `any`: any de construcció de la vivenda.
- `carrer`: tipus de carrer que dona accés a l'inmoble.
- `tipus`: tipus d'habitatge.
- `garatge.cotxes`: número de cotxes que hi caben en el garatge.
- `garatge.area`: àrea del garatge (en peus quadrats).
- `aire`: variable que indica si la vivenda té aire condicionat (Y si en té, N si no).

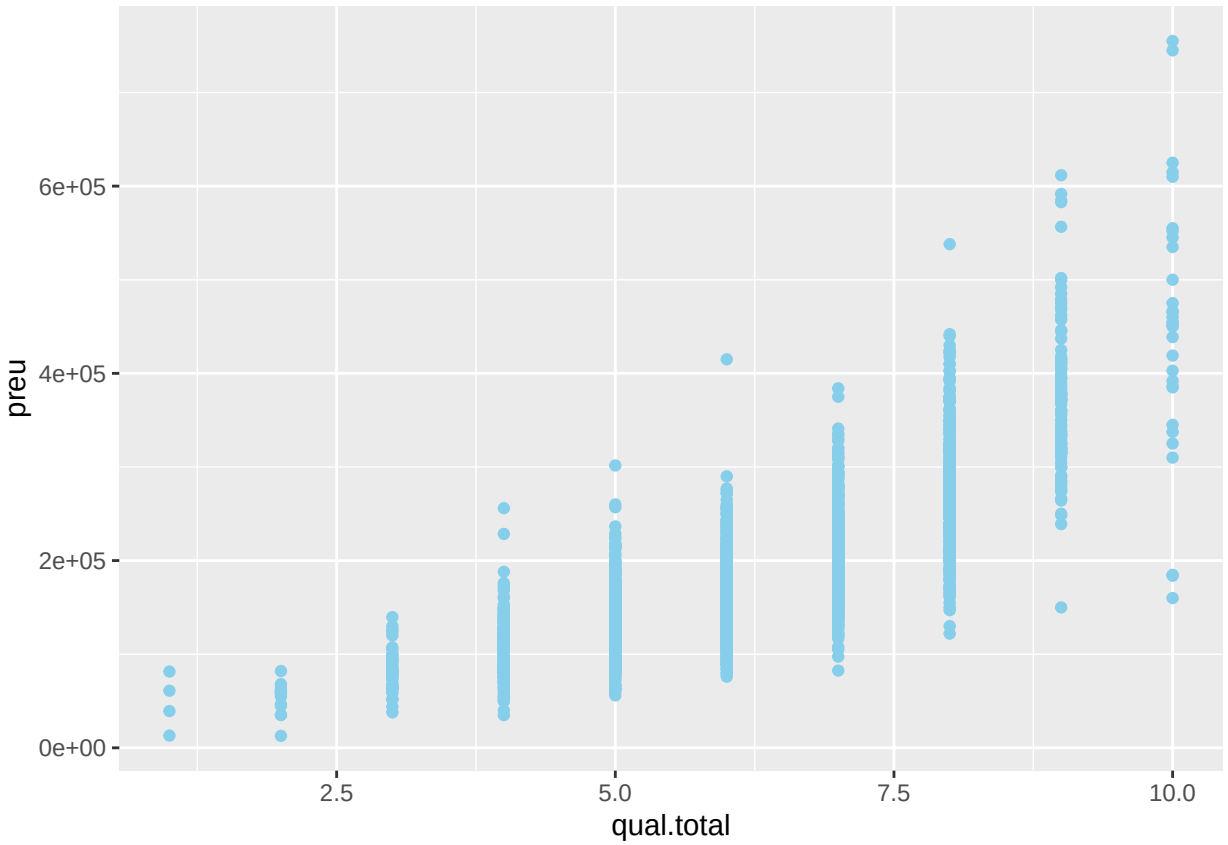
Observeu que tenim cinc variables numèriques (`preu`, `area`, `any`, `garatge.area`, `garatge.cotxes`), i quatre variables cinc variables categòriques (`qual.total`, `condicio.total`, `carrer`, `tipus`, `aire`).

Fem una exploració de les dades a través de diferents gràfics que ens ajudaran a saber com es comporten les nostres variables amb respecte la variable que volem predir (`preu`).

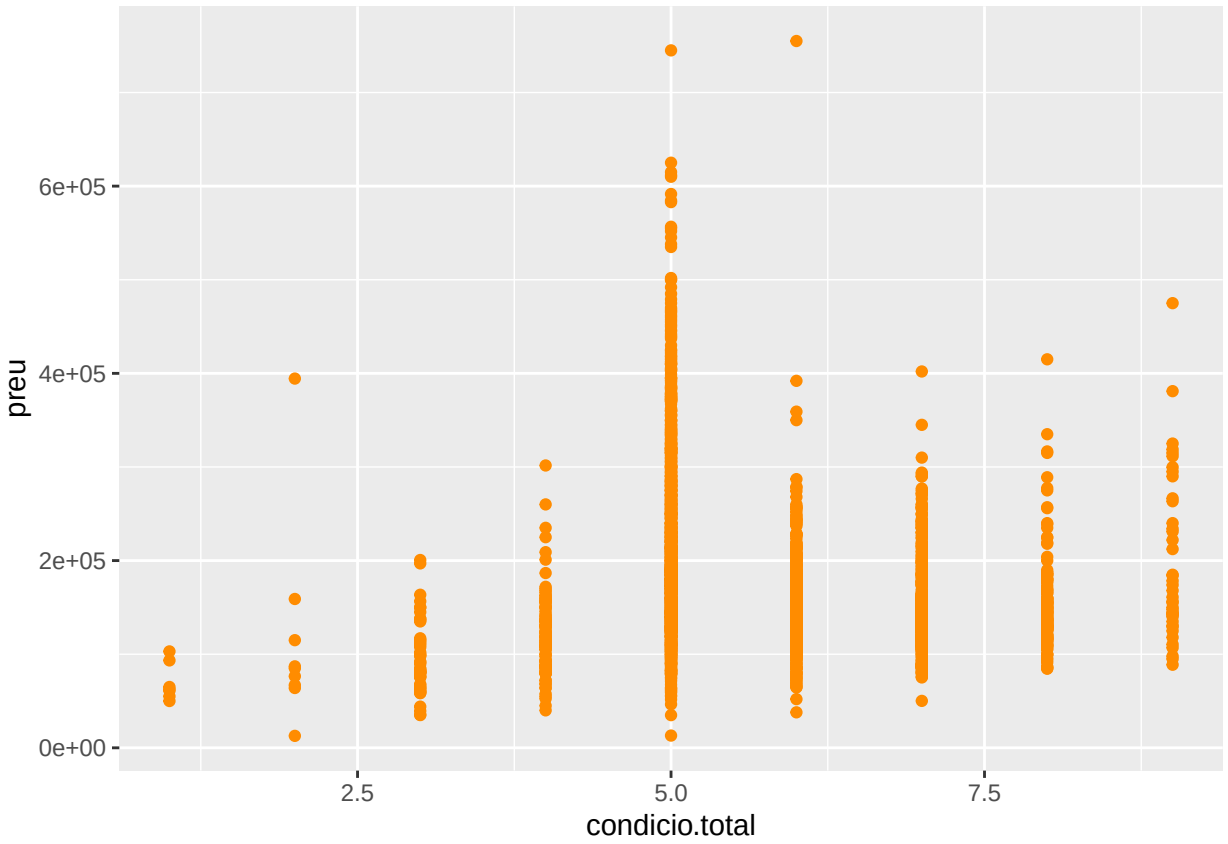
```
library(ggplot2)
ggplot(data = cases_train) +
  geom_point(aes(x = area, y = preu), color = "tomato")
```



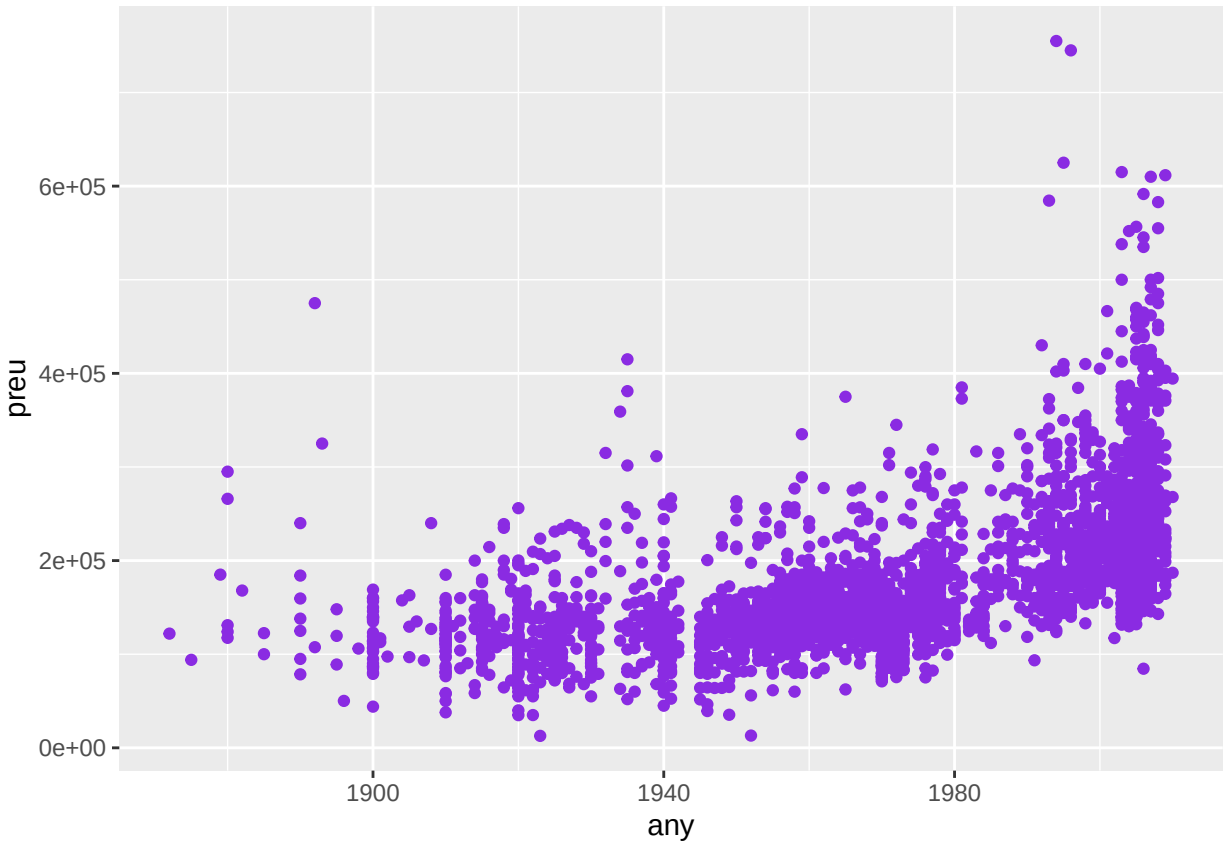

```
ggplot(data = cases_train) +
  geom_point(aes(x = qual.total, y = preu), color = "skyblue")
```



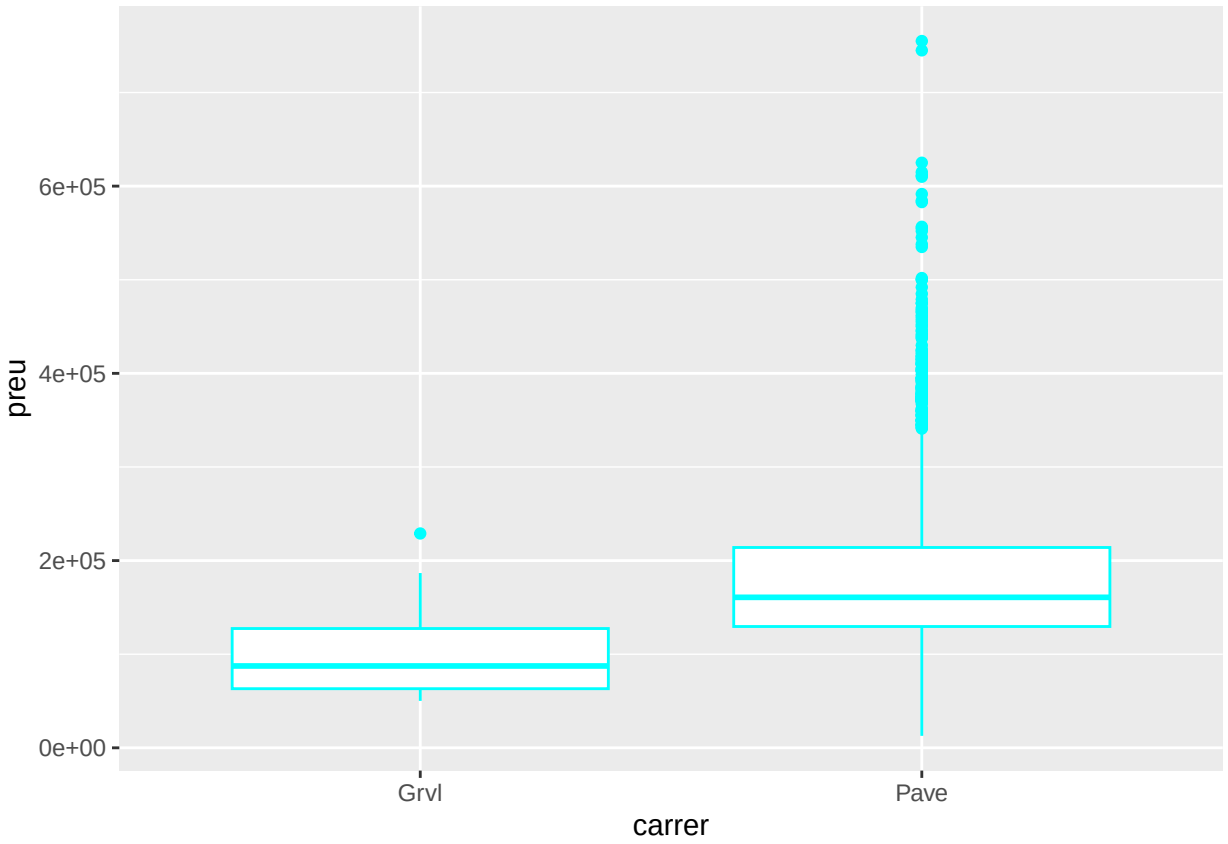

```
ggplot(data = cases_train) +
  geom_point(aes(x = condicio.total, y = preu), color = "darkorange")
```



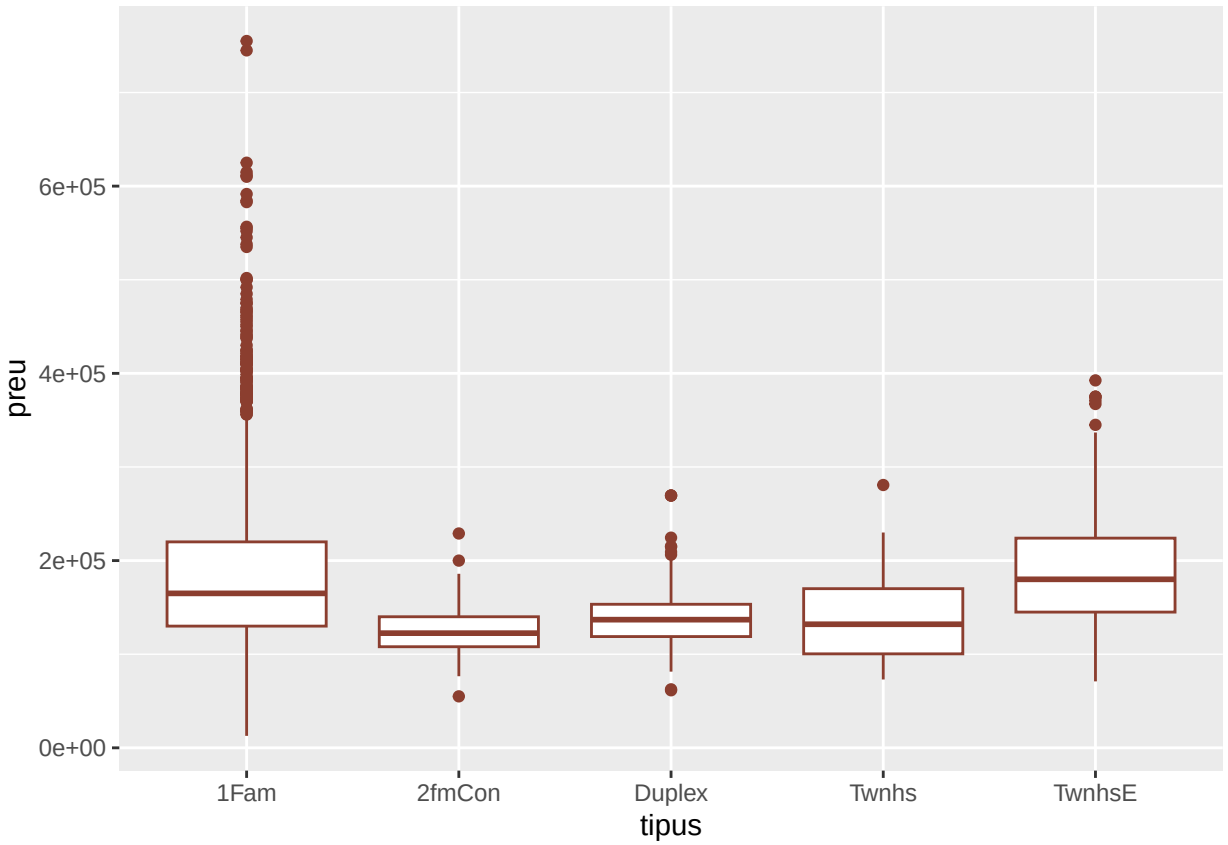

```
ggplot(data = cases_train) +
  geom_point(aes(x = any, y = preu), color = "blueviolet")
```




```
ggplot(data = cases_train) +
  geom_boxplot(aes(x = carrer, y = preu), color = "cyan")
```

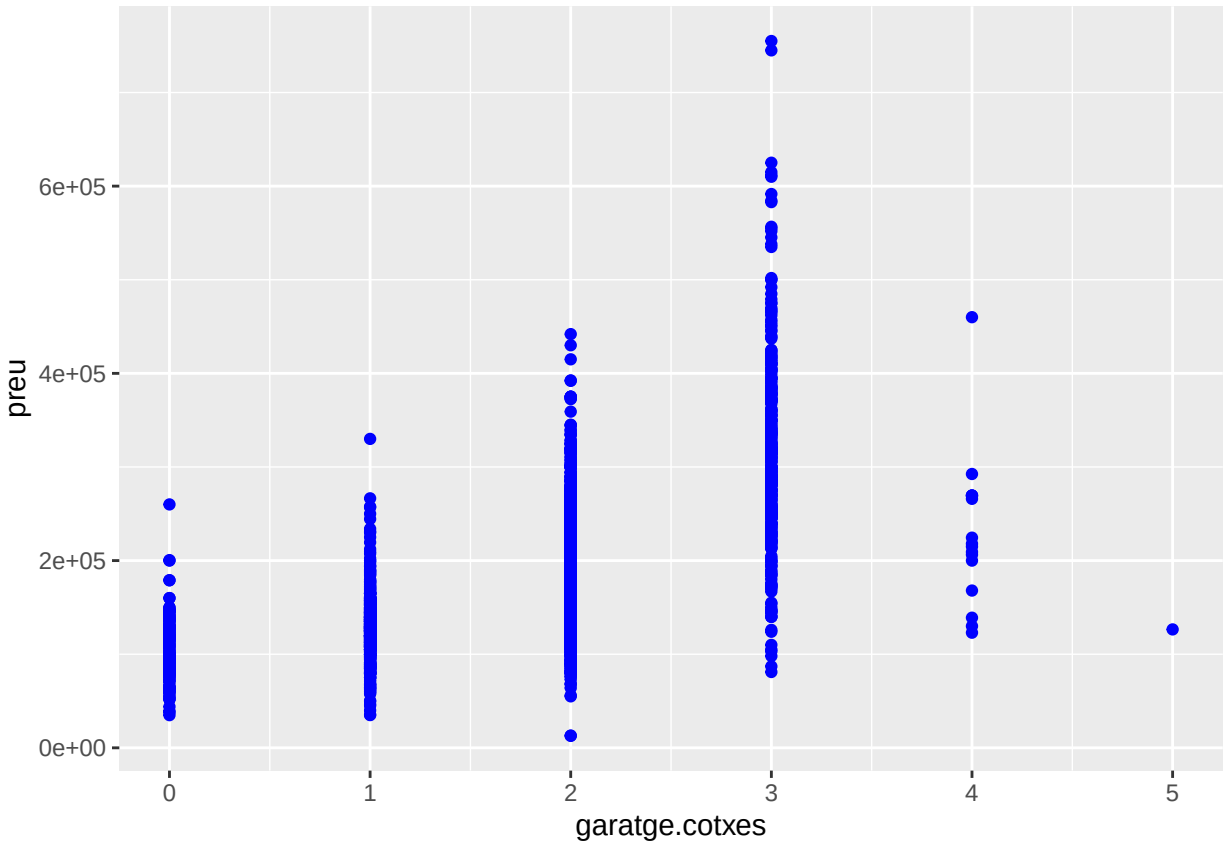



```
ggplot(data = cases_train) +
  geom_boxplot(aes(x = tipus, y = preu), color = "coral4")
```



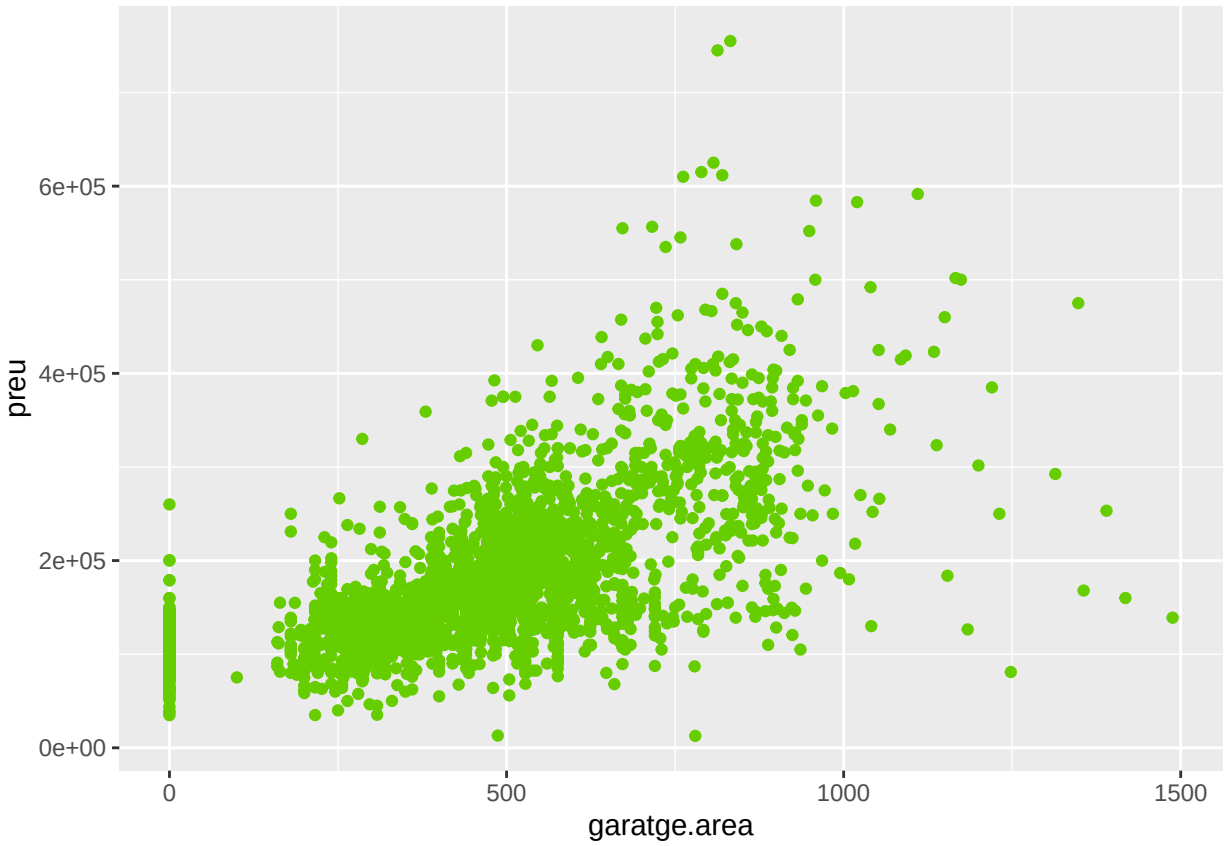

```
ggplot(data = cases_train) +
  geom_point(aes(x = garatge.cotxes, y = preu), color = "blue")
```

```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```

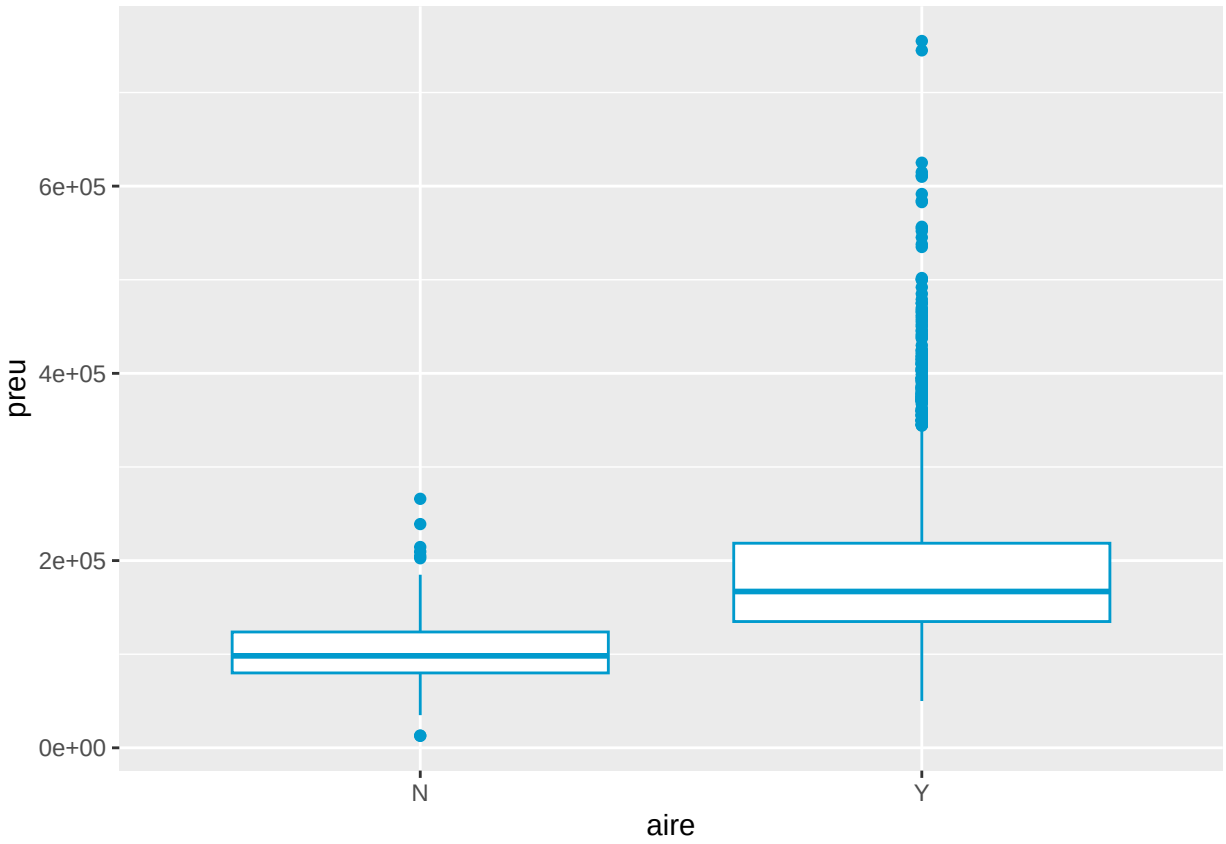



```
ggplot(data = cases_train) +
  geom_point(aes(x = garatge.area, y = preu), color = "chartreuse3")
```

```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```




```
ggplot(data = cases_train) +
  geom_boxplot(aes(x = aire, y = preu), color = "deepskyblue3")
```



Conclusions sobre els gràfics:

- Com més àrea té, més car l'habitatge.
- Com més qualitat total té l'inmoble, més car.
- Sembla que en general, com millor condició general, més car.
- Sembla que la tendència respecte l'any de construcció és: més nou, més car.
- Els habitatges amb carrer tipus `Pave` son més cars que els tipus `Grvl`.
- Sembla que les cases unifamiliars són més cares que les altres.
- En general, sembla que com més cotxes hi caben en el garatge, més car l'habitatge.
- Com més àrea té el garatge, més car l'habitatge.
- Sembla que aquells habitatges amb aire condicionat són més cars.

Ajustem un model lineal (podeu veure informació sobre el model lineal a https://es.wikipedia.org/wiki/Regresi%C3%B3n_lineal). Ajustarem el següent model lineal:

$$preu_i = \beta_1 area_i + \beta_2 qual.total_i + \beta_3 condicio.total_i + \beta_4 any_i + \beta_5 tipus_i + \beta_6 garatge.area_i + \beta_7 aire_i$$

El codi per fer-ho és:

```
model.lineal <- lm(preu ~ area + qual.total + condicio.total + any + tipus +
          garatge.area + aire, data = cases_train)
summary(model.lineal)
```

```
##
## Call:
## lm(formula = preu ~ area + qual.total + condicio.total + any +
##     tipus + garatge.area + aire, data = cases_train)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -419189  -21158   -3136   15775  287734
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.188e+06  6.931e+04 -17.133  < 2e-16 ***
## area            5.772e+01  1.885e+00  30.620  < 2e-16 ***
## qual.total      2.214e+04  8.060e+02  27.475  < 2e-16 ***
## condicio.total  4.978e+03  7.379e+02   6.746 1.83e-11 ***
## any             5.599e+02  3.572e+01  15.675  < 2e-16 ***
## tipus2fmCon    -8.954e+03  5.128e+03  -1.746   0.0809 .
## tipusDuplex    -2.543e+04  3.816e+03  -6.663 3.21e-11 ***
## tipusTwnhs     -3.255e+04  3.988e+03  -8.163 4.86e-16 ***
## tipusTwnhsE    -1.141e+04  2.784e+03  -4.098 4.29e-05 ***
## garatge.area    5.551e+01  4.306e+00  12.893  < 2e-16 ***
## aireY          -8.160e+03  3.264e+03  -2.500   0.0125 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37530 on 2830 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.782,  Adjusted R-squared:  0.7812
## F-statistic:  1015 on 10 and 2830 DF,  p-value: < 2.2e-16
```

Analitzem els signes dels coeficients (columna "Estimate") que hem obtingut en el model lineal:

- `area`: positiu.
- `qual.total`: positiu.
- `condicio.total`: positiu.
- `any`: positiu.
- `tipus`: els signes de tots els coeficients del tipus de casa són negatius amb respecte el tipus `1fam`, el que vol dir que per aquelles cases que no són de tipus unifamiliar el preu és més baix.
- `garatge.area`: positiu.
- `aire`: el signe per la categoria `Y` és negatiu, el que vol dir que per aquells habitatges amb aire condicionat el preu és més barat.

Per a totes les variables numèriques amb signe positiu, el que interpretem és que quan augmenten, també ho fan els preus dels habitatges. Observem que els coeficients obtinguts encaixen amb les anteriors interpretacions pels gràfics, menys en el cas de l'aire condicionat.

Ara, descarreguem una taula de dades que teniem guardada apart. En aquesta taula de dades hi ha dades noves sobre el problema que estem tractant, aquestes dades no s'han fet servir per entrenar el nostre model. El que farem amb aquestes noves dades és testejar el nostre model i veure si és capaç de fer bones prediccions.

Llegim les noves dades:

```
cases_test <- read.csv("cases_test.csv")
cases_test$X <- NULL
head(cases_test)
```

```
##     preu area qual.total condicio.total  any carrer  tipus garatge.cotxes
## 1 127000 1040          5              8 1962   Pave   1Fam              1
## 2 169000 1494          6              6 1974   Pave   1Fam              2
## 3 132000 1268          5              6 1954   Pave   1Fam              1
## 4 143750  848          6              5 2003   Pave TwnhsE              2
## 5 126000  856          6              6 1939   Pave   1Fam              2
```

```
## 6 100000 1666          5              2 1931    Pave    1Fam              0
##   garatge.area aire
## 1          260    Y
## 2          461    Y
## 3          244    Y
## 4          420    Y
## 5          399    Y
## 6            0    Y
```

En aquestes noves dades tenim informació sobre 88 habitatges d'Ames. A continuació, fem prediccions amb el model que hem ajustat anteriorment, i veiem si aquestes prediccions s'acosten al valor real de les noves dades:

```
preds <- predict(model.lineal, newdata = cases_test)
comparison <- data.frame(Real = cases_test$preu, Predit = preds,
                         Difer = abs(cases_test$preu - preds))
head(comparison, 10)
```

```
##      Real   Predit      Difer
## 1  127000 127750.6   750.5647
## 2  169000 184018.8 15018.8123
## 3  132000 125586.4  6413.5833
## 4  143750 144307.0   557.0225
## 5  126000 124159.1  1840.9353
## 6  100000 102222.6  2222.5760
## 7  185000 206353.6 21353.6221
## 8  169000 142301.8 26698.1670
## 9  149900 136187.8 13712.1759
## 10 254000 224966.0 29034.0002
```

Només estem ensenyant 10 resultats de les 88 prediccions que hem fet. Com podeu veure el nostre model només s'equivoca predint el preu real de l'habitatge per 750 dòlars (en una casa que costa 127mil dòlars), 15mil dòlars (en una casa que costa 169mil dòlars), 6mil dòlars (en una casa que costa 132mil dòlars), 557 dòlars (en una casa que costa 143mil dòlars), etc.

Predicció d'espècies de pingüins

En aquest segon model volem predir espècies de pingüins tenint en compte característiques físiques d'aquests. Les tres espècies de pinguins amb les que treballarem són: Gentoo, Adelie i Chinstrap. A continuació llegim la base de dades i imprimim algunes de les observacions de la taula de dades que farem servir:

```
penguins_train <- read.csv("penguins_train.csv")
penguins_train$X <- NULL
# convertim en factors les variables categòriques
penguins_train$sex <- factor(penguins_train$sex)
head(penguins_train)
```

```
##     species    island bill_length_mm bill_depth_mm flipper_length_mm
## 1    Gentoo    Biscoe           49.5          16.1               224
## 2    Gentoo    Biscoe           55.1          16.0               230
## 3    Adelie     Dream           39.7          17.9               193
## 4    Adelie    Biscoe           35.0          17.9               190
## 5    Adelie Torgersen           40.6          19.0               199
## 6 Chinstrap     Dream           52.0          18.1               201
##   body_mass_g    sex
## 1        5650   male
## 2        5850   male
## 3        4250   male
```

```
## 4        3450 female
## 5        4000   male
## 6        4050   male
```

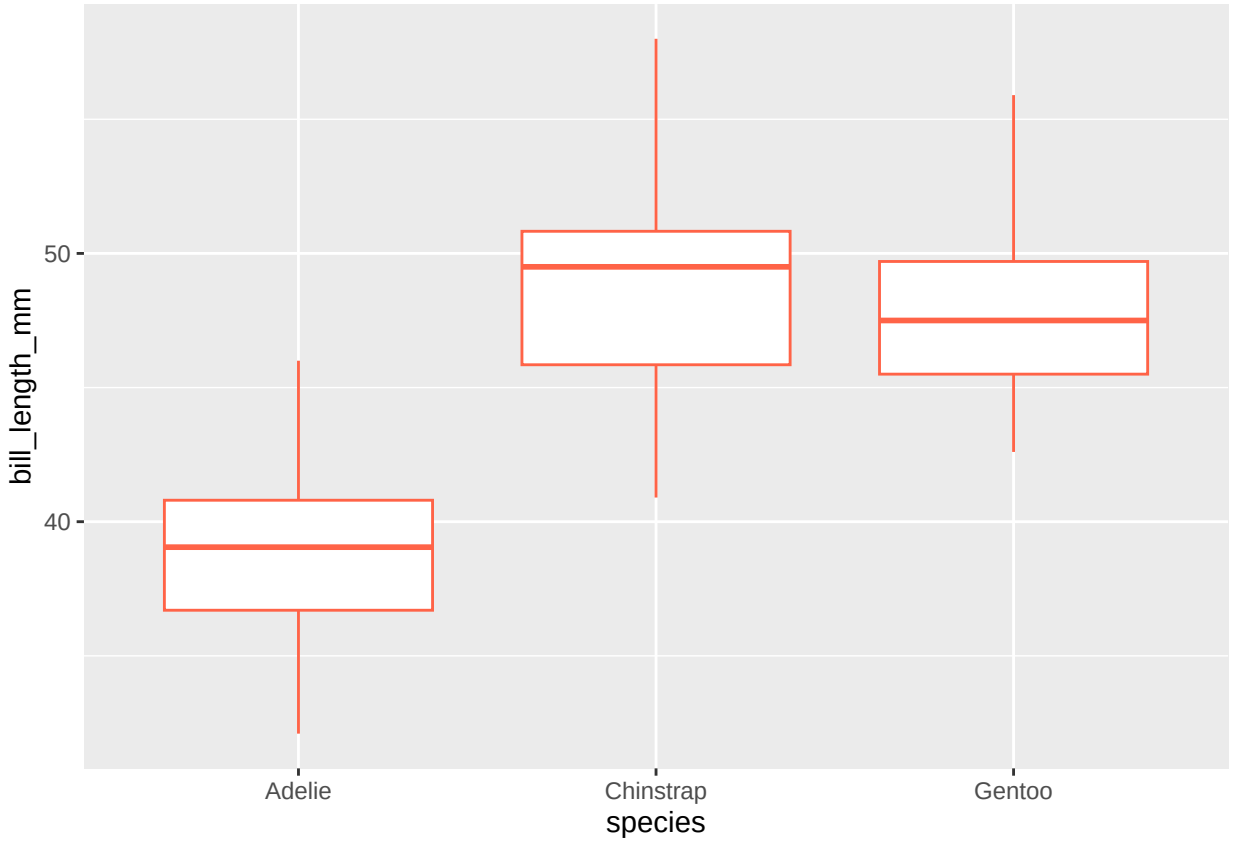
En aquesta taula de dades que farem servir per entrenar el model predictiu tenim informació sobre 233 pingüins diferents, i per cadascún d'ells tenim informació sobre 7 característiques. Les variables que tenim en aquest cas són:

- `species`: espècie del pingüi (pot ser Gentoo, Adelie o Chinstrap).
- `island`: illa de procedència del pingüi.
- `bill_length_mm`: longitud del bec (en mm).
- `bill_depth_mm`: profunditat del bec (en mm).
- `flipper_length_mm`: longitud de les aletes (en mm).
- `body_mass_g`: pes (en grams).
- `sex`: sexe del pingüi.

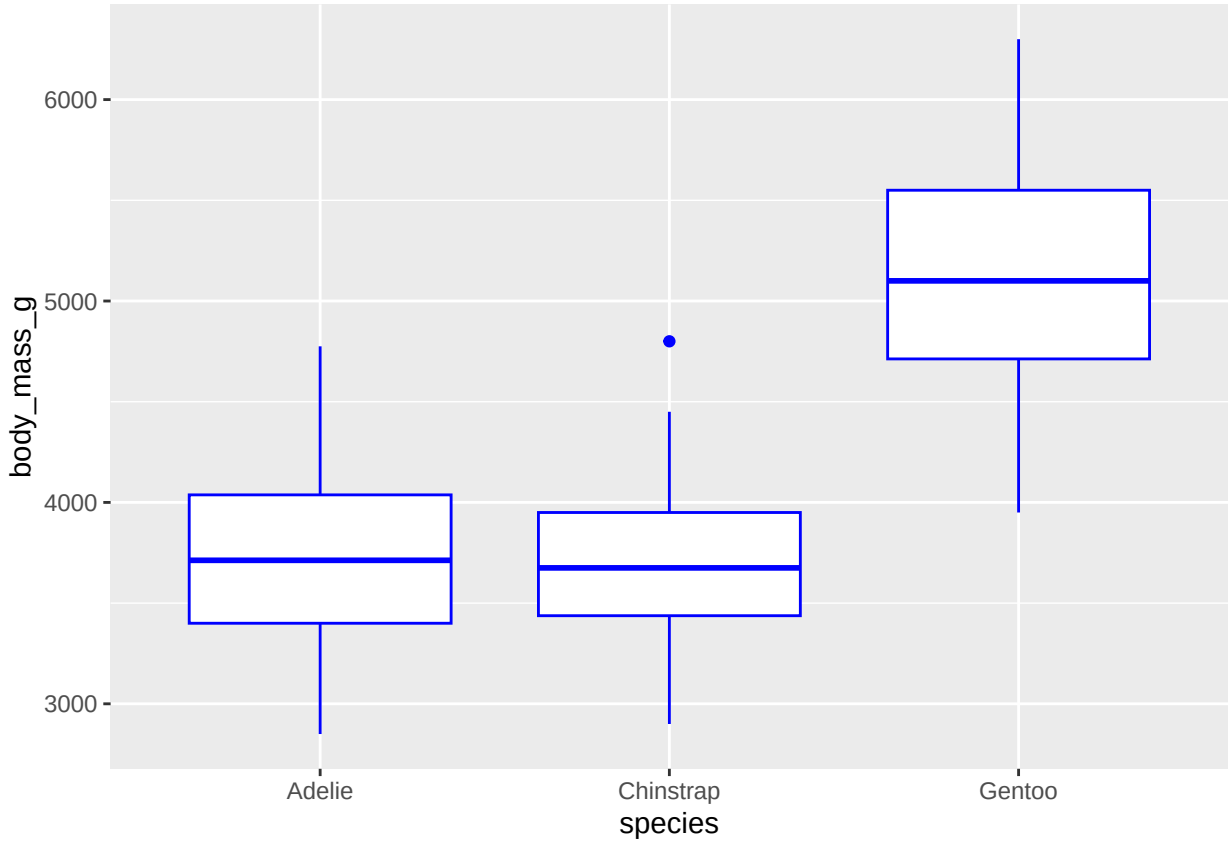
Ara tenim quatre variable numèriques (`bill_length_mm`, `bill_depth_mm`, `flipper_length_mm`, `body_mass_g`) i tres variables categòriques (`species`, `island`, `sex`). Una altra observació important, és que la variable que volem predir en aquest cas (`species`) és categòrica. Mentre que abans volíem predir una variable numèrica i contínua (el preu d'habitatge), ara volem predir una variable categòrica. Així doncs, haurem de fer servir un altre model diferent. En aquest cas farem servir un model lineal generalitzat (veure més informació a https://es.wikipedia.org/wiki/Modelo_lineal_generalizado).

Primerament, tal i com hem fet abans, explorem les dades que tenim fent servir alguns gràfics:

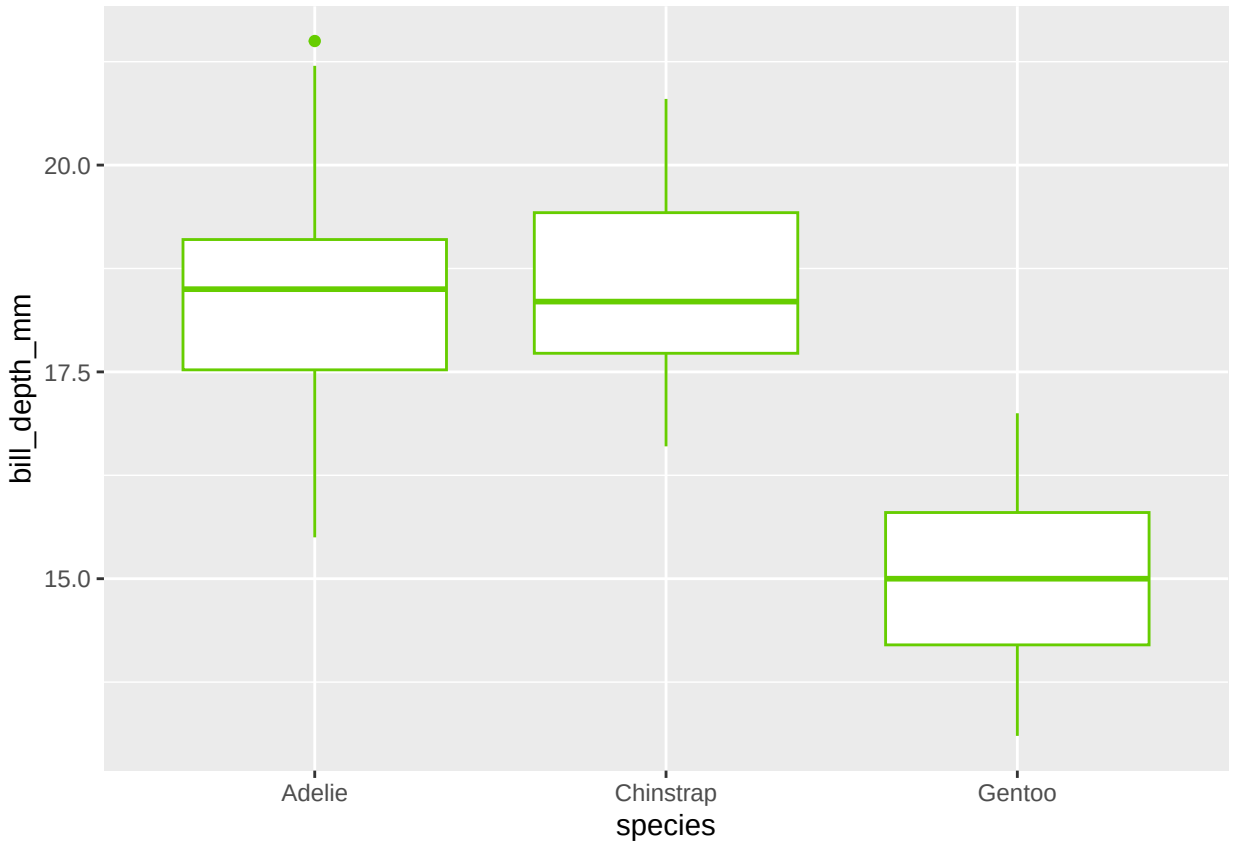
```
ggplot(penguins_train) +
  geom_boxplot(aes(x=species, y=bill_length_mm), color = "tomato")
```



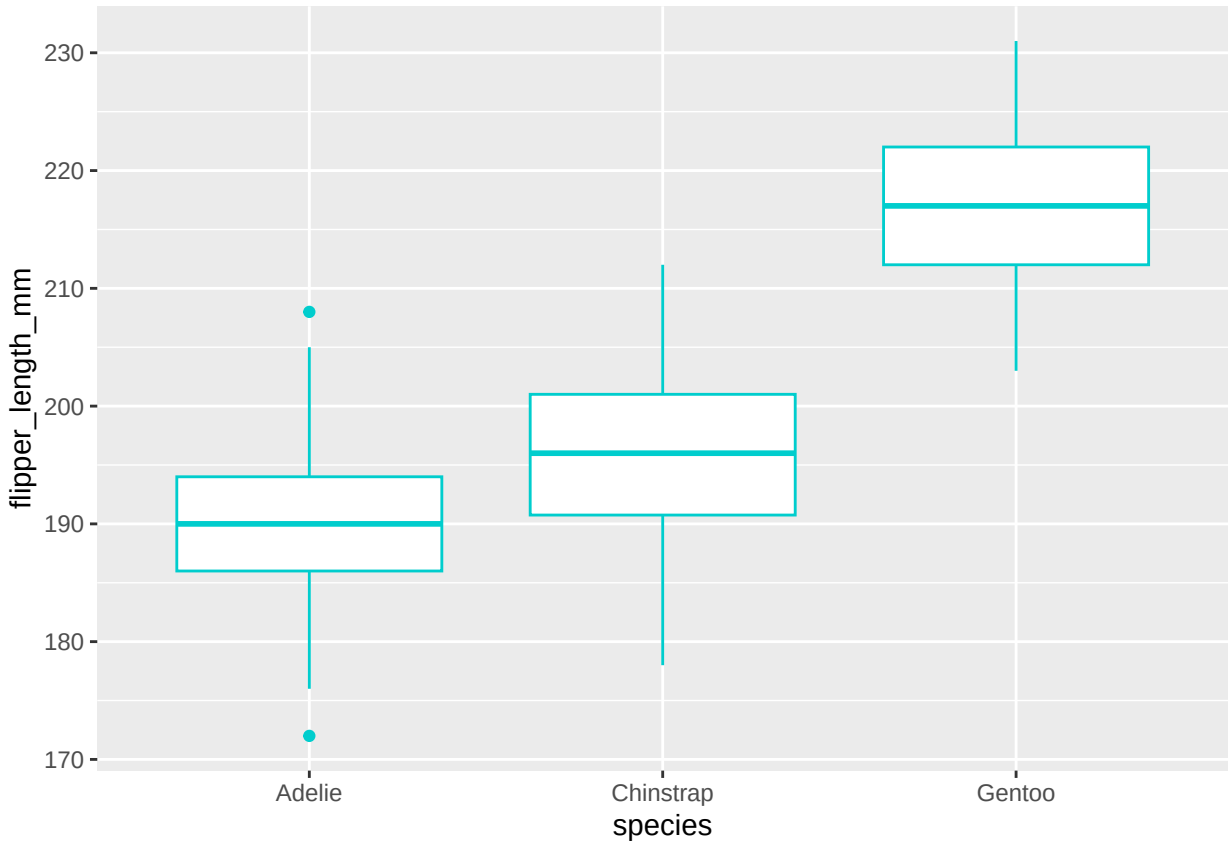

```
ggplot(penguins_train)+
  geom_boxplot(aes(x=species, body_mass_g), color = "blue")
```



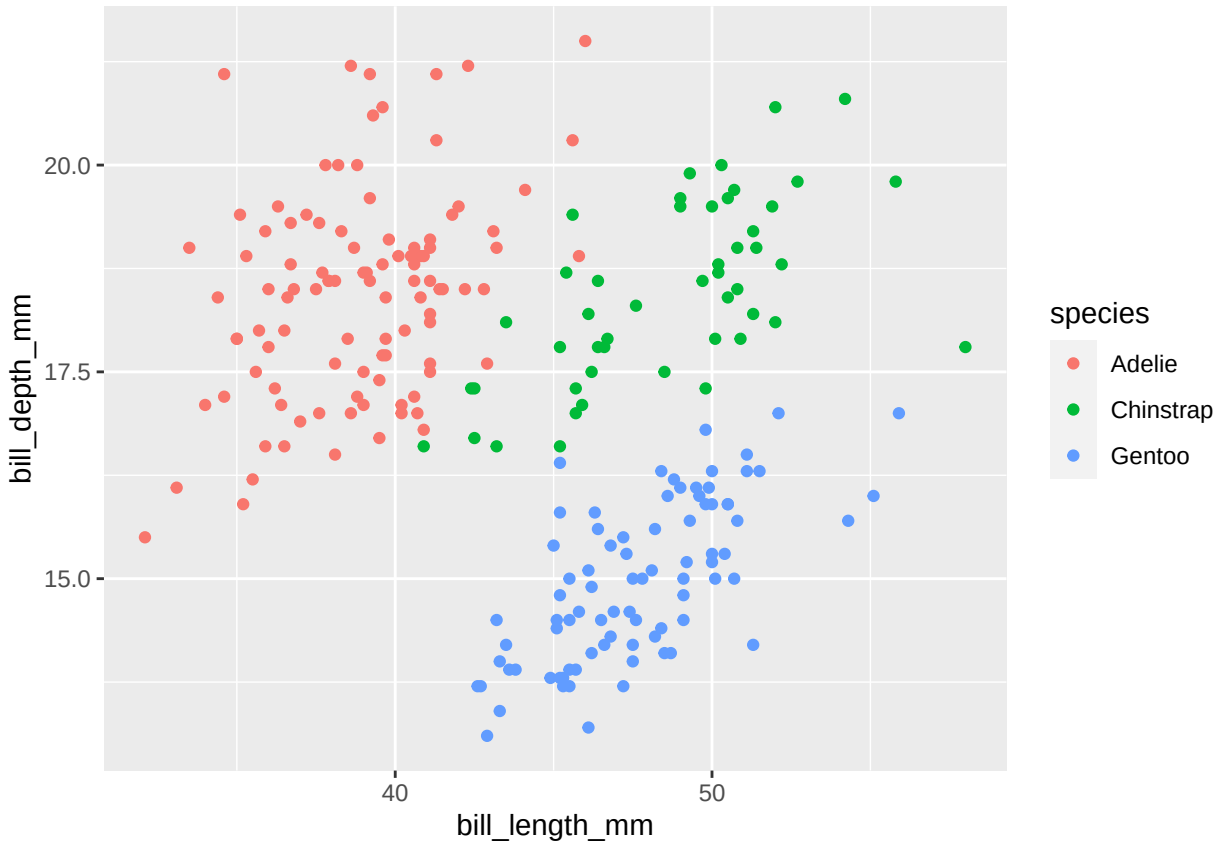

```
ggplot(penguins_train)+
  geom_boxplot(aes(x=species, y=bill_depth_mm), color = "chartreuse3")
```




```
ggplot(penguins_train)+
  geom_boxplot(aes(x=species, y=flipper_length_mm), color = "cyan3")
```




```
ggplot(penguins_train)+
  geom_point(aes(x=bill_length_mm, y=bill_depth_mm, color=species))
```



Interpretem els gràfics:

- Sembla que els pingüins Chinstrap tenen el bec més llarg que els Gentoo, i aquests més que els Adalie.
- Sembla que els pingüins Gentoo pesen més que els altres.
- Sembla que els Gentoo tenen un bec menys profund.
- Sembla que els Gentoo tenen aletes més llargues que els Chinstrap, i aquests més llargues que els Adelie.
- En el darrer gràfic sembla que es por veure un patró que separa les tres espècies. Podrà el nostre model captar aquesta separació?

També podem fer una taula per veure quines espècies de pingüins hi ha a cada illa:

```
with(penguins_train, table(species, island))
```

```
##            island
## species     Biscoe Dream Torgersen
##   Adelie        29    40        33
##   Chinstrap      0    48         0
##   Gentoo        83     0         0
```

Veiem que a la illa de Biscoe hi ha 29 Adelies, cap Chinstrap i 83 Gentoos. A la illa de Dream hi ha 40 Adelies, 48 Chinstraps i cap Gentoo. I a la illa de Torgersen només hi ha 33 Adalies.

Ajustem ara el model lineal generalitzat:

```
library(nnet)
model.lingen <- step(multinom(species ~ . , data = penguins_train,
                              trace = FALSE), trace = 0)
```

```
## trying - island
## trying - bill_length_mm
```

```
## trying - bill_depth_mm
## trying - flipper_length_mm
## trying - body_mass_g
## trying - sex
## trying - bill_length_mm
## trying - bill_depth_mm
## trying - flipper_length_mm
## trying - body_mass_g
## trying - sex
## trying - bill_length_mm
## trying - bill_depth_mm
## trying - flipper_length_mm
## trying - body_mass_g
## trying - bill_length_mm
## trying - bill_depth_mm
## trying - body_mass_g
```

```
summary(model.lingen)
```

```
## Call:
## multinom(formula = species ~ bill_length_mm + bill_depth_mm +
##     body_mass_g, data = penguins_train, trace = FALSE)
##
## Coefficients:
##           (Intercept) bill_length_mm bill_depth_mm body_mass_g
## Chinstrap  -31.421711       26.67539     -45.06581 -0.08653157
## Gentoo       1.831695       18.33250     -50.30234  0.01703379
##
## Std. Errors:
##            (Intercept) bill_length_mm bill_depth_mm body_mass_g
## Chinstrap 1.414604e-01   3.3218662775   1.933656306  0.05243895
## Gentoo    8.720521e-05   0.0004517812   0.002566747  0.02930422
##
## Residual Deviance: 0.02044068
## AIC: 16.02044
```

Ara que ja tenim el model predictiu entrenat, ens baixem unes noves dades per comprovar la capacitat de predicció del nostre model, tal i com hem fet abans.

A continuació ens baixem les dades per fer el test:

```
penguins_test <- read.csv("penguins_test.csv")
penguins_test$year <- NULL
penguins_test$sex <- factor(penguins_test$sex)
head(penguins_test)
```

```
##   species    island bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
## 1  Adelie Torgersen           38.9          17.8               181        3625
## 2  Adelie Torgersen           36.6          17.8               185        3700
## 3  Adelie Torgersen           42.5          20.7               197        4500
## 4  Adelie    Biscoe           37.8          18.3               174        3400
## 5  Adelie    Biscoe           38.2          18.1               185        3950
## 6  Adelie    Biscoe           40.5          17.9               187        3200
##      sex
## 1 female
## 2 female
```

```
## 3   male
## 4 female
## 5   male
## 6 female
```

En aquestes dades de test tenim informació sobre 100 pingüins.

El que farem ara és dur a terme les prediccions de les noves dades fent servir el model que hem entrenat anteriorment. Un cop fetes les prediccions construirem l'anomenada matriu de confusió. En aquesta matriu podrem veure les espècies que ha predit el nostre model, juntament amb les espècies que són correctes (les que hi ha a la nova taula de dades que estem fent servir).

```
library(caret)
```

```
## Loading required package: lattice
```

```
preds_train <- predict(model.lingen, newdata = penguins_test)
confusionMatrix(preds_train, factor(penguins_test$species))
```

```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction  Adelie Chinstrap Gentoo
##   Adelie        44         1      0
##   Chinstrap      0        19      0
##   Gentoo         0         0     36
##
## Overall Statistics
##
##                Accuracy : 0.99
##                  95% CI : (0.9455, 0.9997)
##     No Information Rate : 0.44
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9842
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: Adelie Class: Chinstrap Class: Gentoo
## Sensitivity                 1.0000           0.9500          1.00
## Specificity                 0.9821           1.0000          1.00
## Pos Pred Value              0.9778           1.0000          1.00
## Neg Pred Value              1.0000           0.9877          1.00
## Prevalence                  0.4400           0.2000          0.36
## Detection Rate              0.4400           0.1900          0.36
## Detection Prevalence        0.4500           0.1900          0.36
## Balanced Accuracy           0.9911           0.9750          1.00
```

Com podeu veure el nostre model només s'ha equivocat una vegada! Un dels pingüins ha estat classificat com a Adelie quan en realitat era Chinstrap. En tots els altres casos la classificació ha estat correcte. Podeu veure que, apart de la matriu de confusió, tenim més informació amunt. Podem veure que el nostre model ha fet una classificació amb una precisió (accuracy) del 99%!