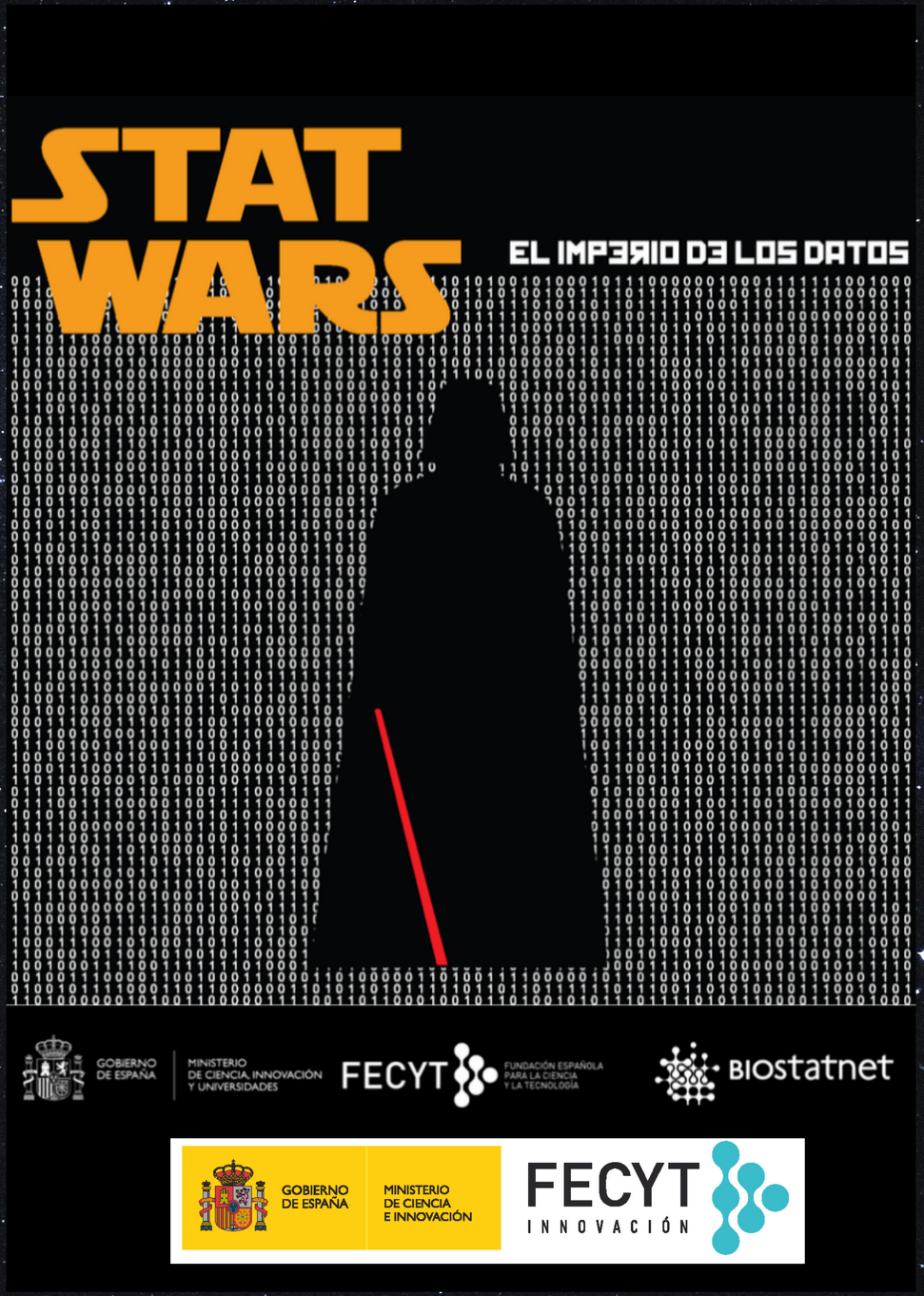
UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

BIOstatnet

GRBIO

El meu primer model predictiu



Arnau G. Fernandez

Introducció

- 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**. Farem servir:
  - Regressions lineals
  - Regressions lineals generalitzades
- Veurem com podem utilitzar les matemàtiques per establir relacions entre diferents variables i fer prediccions amb aquestes.

Introducció

- **R**, és el principal llenguatge de programació pels estadístics. Es tracta d'un software lliure, gratuït que es troba a l'abast de tothom.

Introducció

- Veurem dos models predictius:
  - Predicció de **preus d'habitatges**.
  - Predicció d'**espècies de pingüins**.

Predicció de preus d'habitatges

- 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 habitatgez 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.

Predicció de preus d’habitatges

Dramatització

Jo m’encarrego!

Tenim 88 noves cases a la nostra cartera immobiliària. Necessitem predir els seus preus per posar-les en venda ja!

Base de dades

- Fem servir la base de dades de l'empresa per entrenar el nostre model:

	preu	area	qual.total	condicio.total	any	carrer	tipus	garatge.cotxes	garatge.area	aire
1	181000	1604	6	5	1997	Pave	1Fam	2	470	Y
2	124000	1584	5	5	1967	Pave	Duplex	3	792	Y
3	72000	819	5	4	1919	Pave	1Fam	0	0	N
4	137000	1229	6	6	1980	Pave	TwnhsE	2	462	Y
5	124500	864	4	7	1970	Pave	1Fam	2	463	Y
6	228500	2169	8	5	2002	Pave	1Fam	2	647	Y
7	235000	1525	7	5	2000	Pave	1Fam	2	541	Y
8	204000	2299	6	5	1999	Pave	1Fam	2	482	Y
9	146300	1069	6	5	1975	Pave	TwnhsE	2	440	Y
10	350000	2006	8	5	1998	Pave	1Fam	3	938	Y

Variables

- 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.

Variables

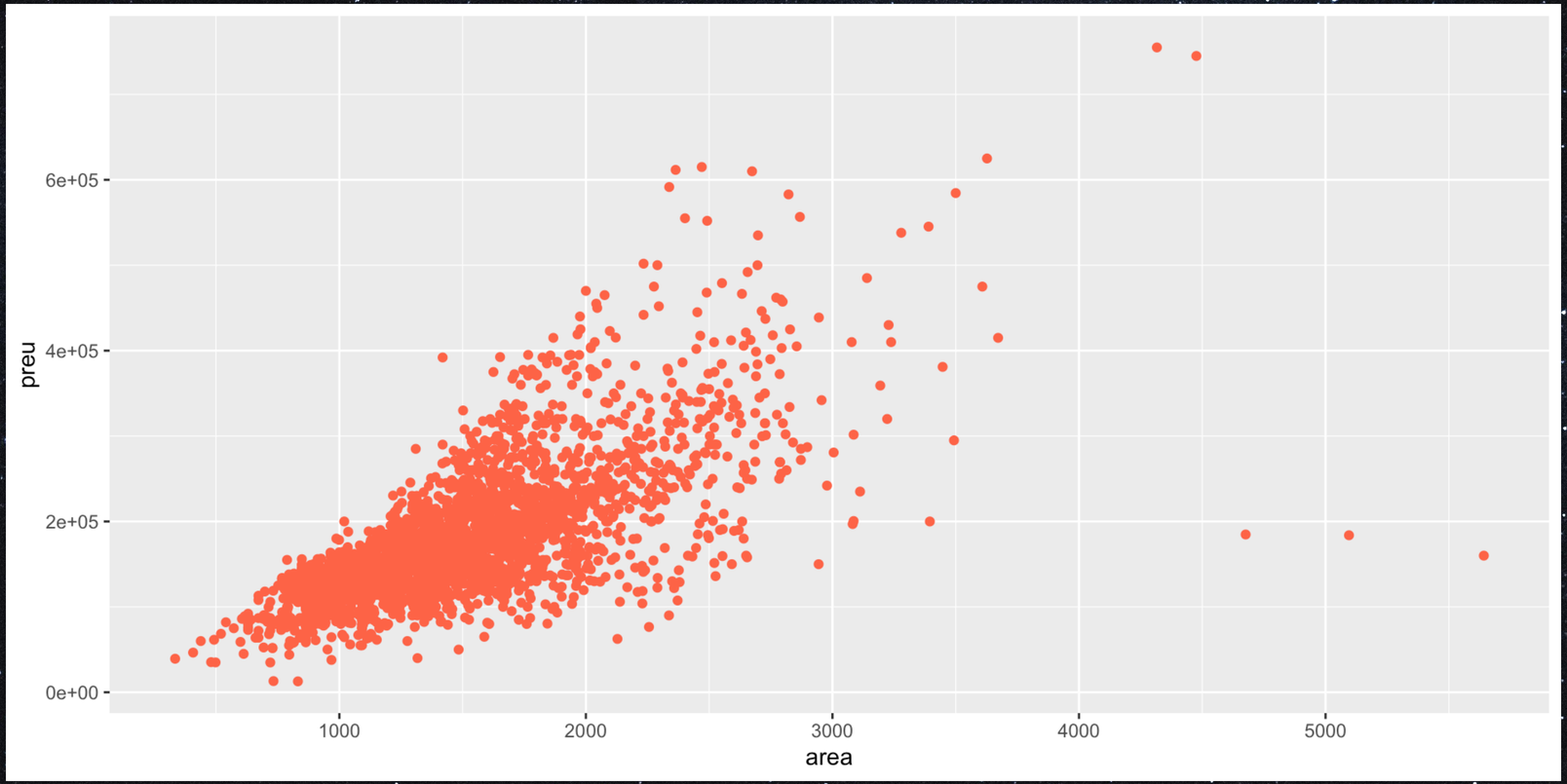
- **carrer**: tipus de carrer que dona accés a l’immoble.
- **tipus**: tipus d'habitatge.
- **garatge.cotxes**: número de cotxes que hi caben en el garatge.
- **garatge.area**: àrea del garatge (en peus quadrats).
- **aire**: variable que indica si la vivenda té aire condicionat (Y si en té, N si no).

Exploració de dades

- 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**).

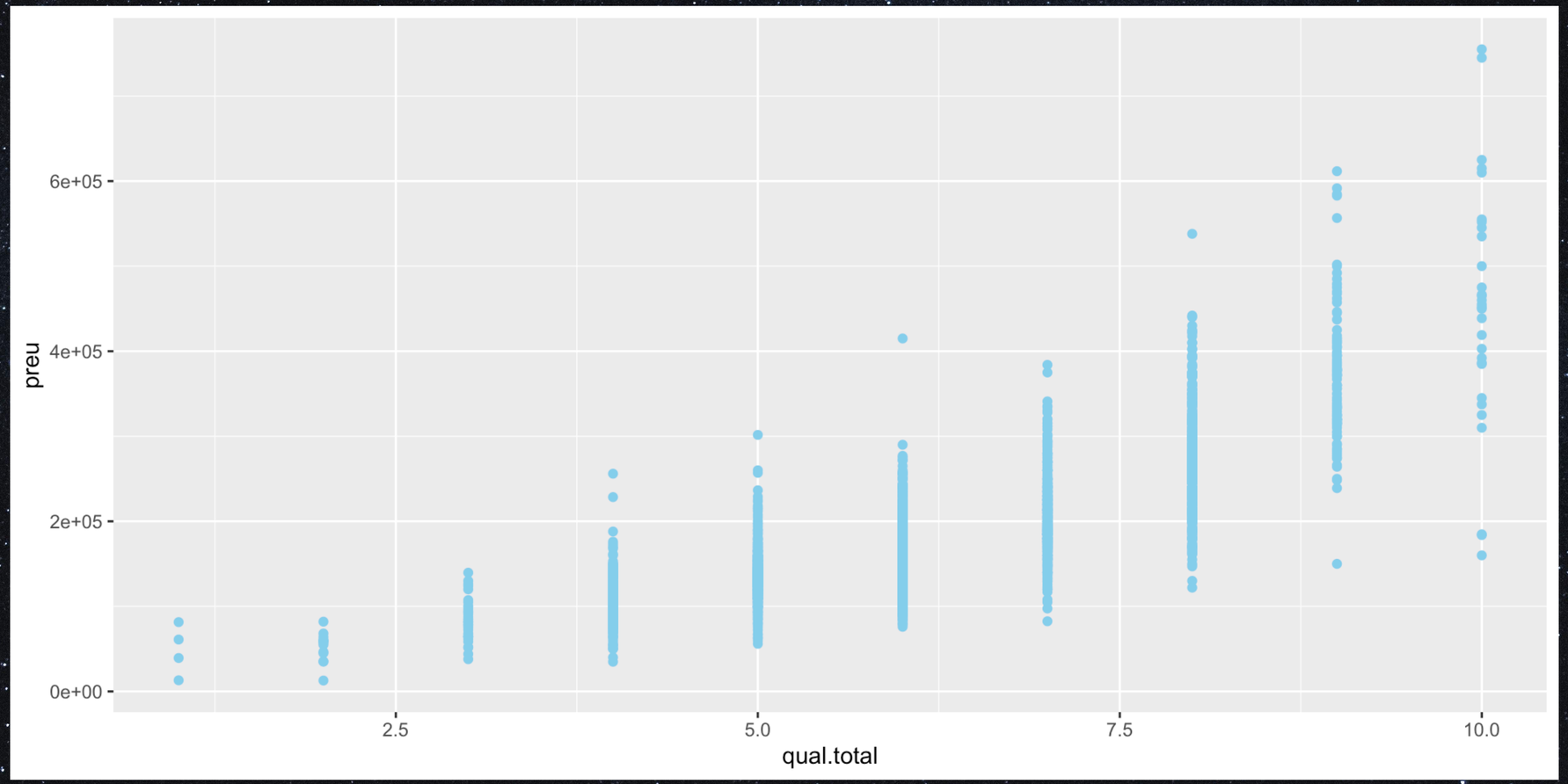
Exploració de dades

Gràfic de punts àrea VS preu



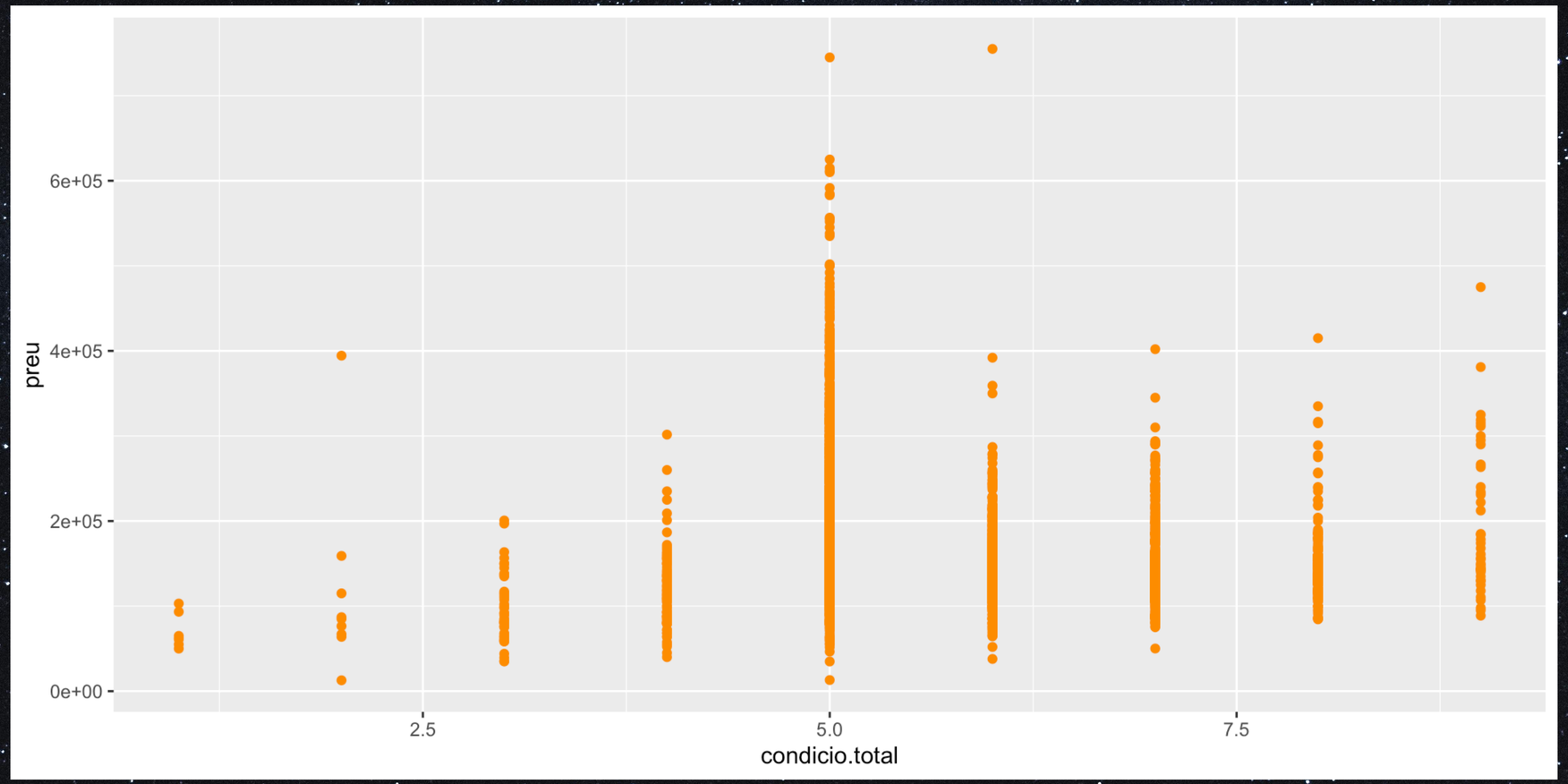
Exploració de dades

Gràfic qualitat total VS preu



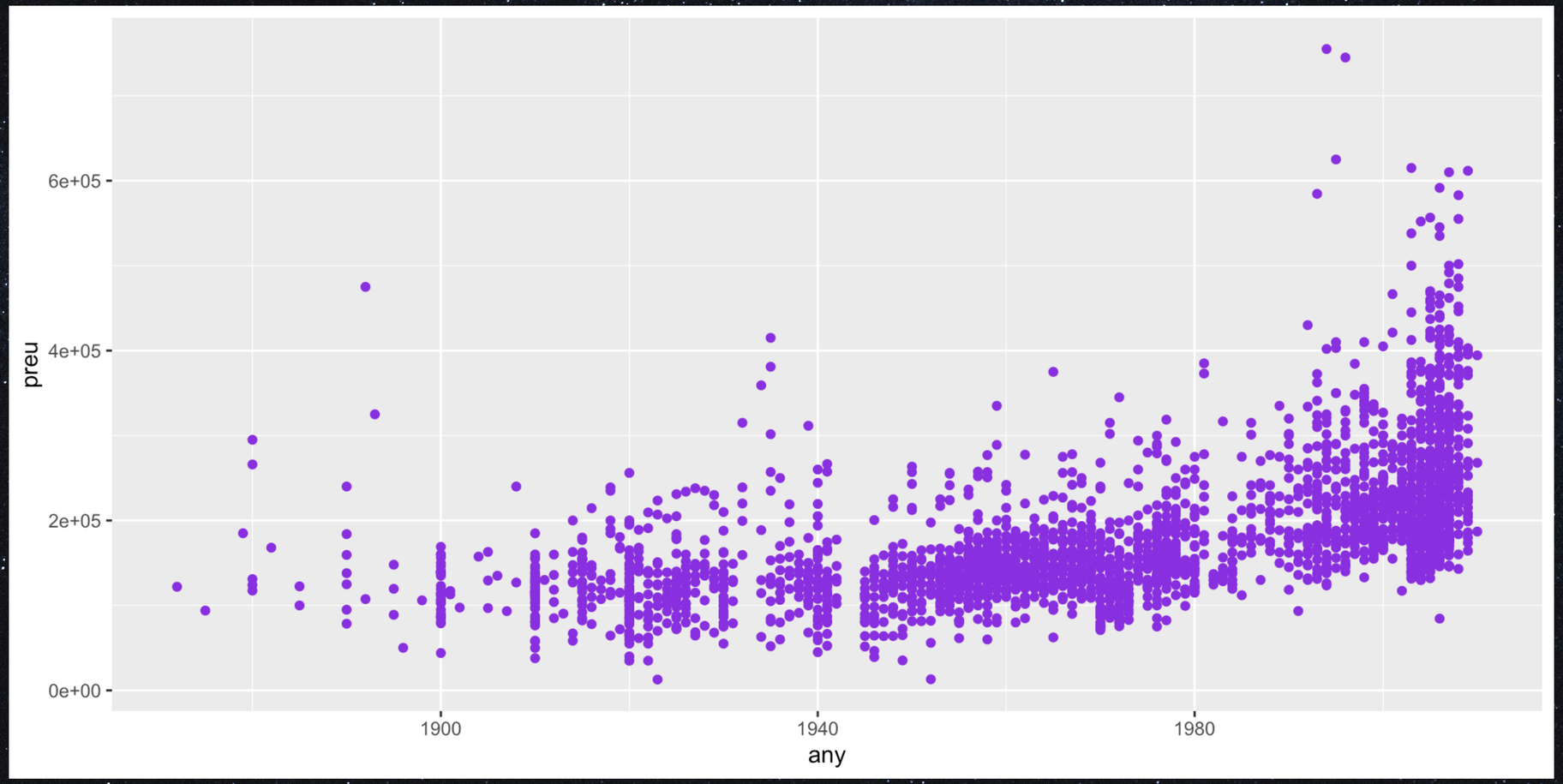
Exploració de dades

Gràfic condició total VS preu



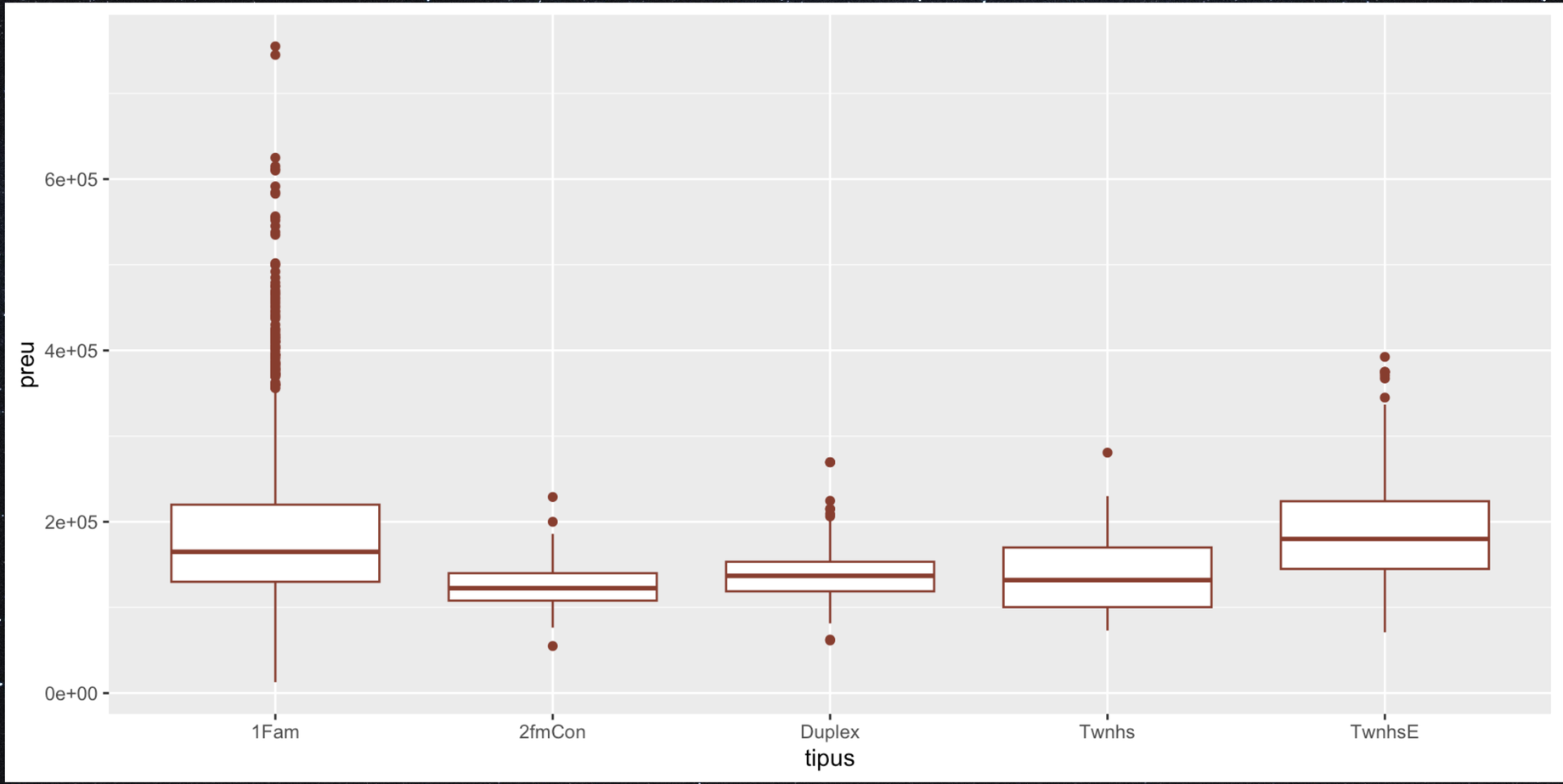
Exploració de dades

Gràfic de punts any VS preu



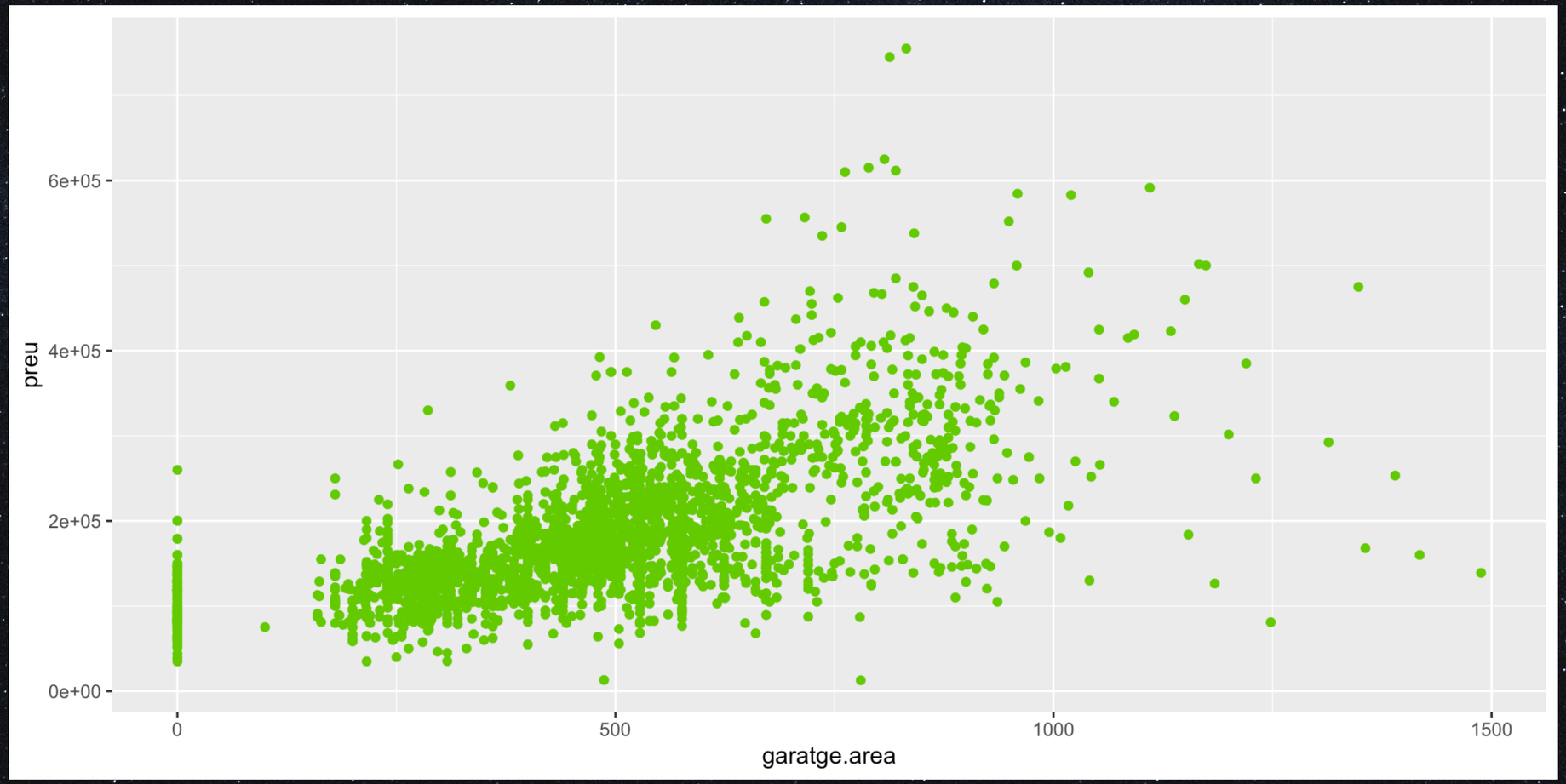
Exploració de dades

Boxplot tipus de case VS preu



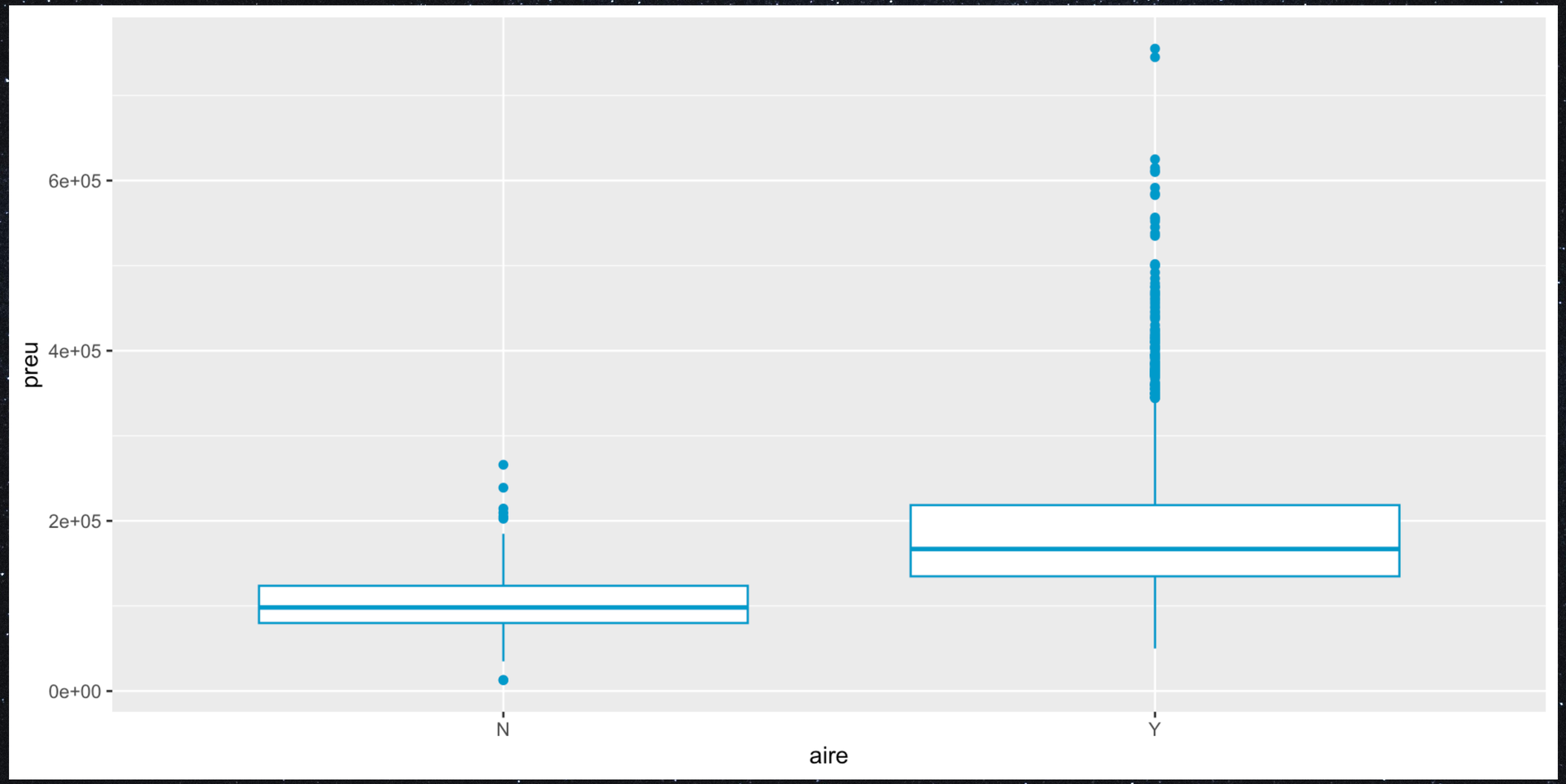
Exploració de dades

Gràfic de punts àrea del garatge VS preu



Exploració de dades

Boxplot aire condicionat VS preu



Model lineal

- Ajustarem el següent model lineal:

$$\texttt{preu}_i = \beta_1\texttt{area}_i + \beta_2\texttt{qual.total}_i + \beta_3\texttt{condicio.total}_i + \beta_4\texttt{any}_i + \beta_5\texttt{tipus}_i + +\beta_6\texttt{garatge.area}_i + \beta_7\texttt{aire}$$

- Els betas són coeficients que cal estimar. El codi per estimar el model a R:

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

Model lineal

Estimació coeficients

```
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 *
---
```

Dades de test

- Fem servir la base de dades que ens ha arribat de les 88 cases noves a la cartera de la immobiliària:

	preu	area	qual.total	condicio.total	any	carrer	tipus	garatge.cotxes	garatge.area	aire
1	127000	1040	5	8	1962	Pave	1Fam	1	260	Y
2	169000	1494	6	6	1974	Pave	1Fam	2	461	Y
3	132000	1268	5	6	1954	Pave	1Fam	1	244	Y
4	143750	848	6	5	2003	Pave	TwnhsE	2	420	Y
5	126000	856	6	6	1939	Pave	1Fam	2	399	Y
6	100000	1666	5	2	1931	Pave	1Fam	0	0	Y
7	185000	1034	8	6	1989	Pave	TwnhsE	2	598	Y
8	169000	1264	5	5	1965	Pave	1Fam	2	528	Y
9	149900	833	5	7	1984	Pave	1Fam	2	495	Y
10	254000	1324	8	5	2006	Pave	TwnhsE	2	550	Y

Prediccions

- A continuació, fem **prediccions** amb el model que hem entrenat anteriorment, i veiem si aquestes prediccions s'acosten al valor real de les noves dades.
- El codi per poder fer prediccions a R:

```
preds <- predict(model.lineal, newdata = cases_test)
```

Prediccions

- Comparem les prediccions del nostre model amb els valors reals:

Real	Predit	Difer
127000	127750.6	750.5647
169000	184018.8	15018.8123
132000	125586.4	6413.5833
143750	144307.0	557.0225
126000	124159.1	1840.9353
100000	102222.6	2222.5760
185000	206353.6	21353.6221
169000	142301.8	26698.1670

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**.

Gentoo

Adelie

Chinstrap

Predicció d'espècies de pingüins

Dramatització

- A aquest biòleg de l'Antàrtida li ha arribat la nova base de dades de pingüins, però els seus companys han oblidat posar l'espècie de pingüí! Podríem fer un model que predigui l'espècie de pingüí tenint en compte les dades?

Base de dades

- Farem servir la base de dades que ja tenim en el grup d'investigació de pingüins per entrenar el nostre model.

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
1	Gentoo	Biscoe	49.5	16.1	224	5650	male
2	Gentoo	Biscoe	55.1	16	230	5850	male
3	Adelie	Dream	39.7	17.9	193	4250	male
4	Adelie	Biscoe	35	17.9	190	3450	female
5	Adelie	Torgersen	40.6	19	199	4000	male
6	Chinstrap	Dream	52	18.1	201	4050	male
7	Gentoo	Biscoe	49.2	15.2	221	6300	male
8	Gentoo	Biscoe	50.1	15	225	5000	male
9	Adelie	Torgersen	37.2	19.4	184	3900	male
10	Gentoo	Biscoe	52.1	17	230	5550	male

Variables

- En aquesta base de dades tenim informació sobre **233 pingüins**. Per cada pingüí tenim informació de **7 variables**.
  - **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 ales (en mm).
  - **body_mass_g**: pes (en grams).
  - **sex**: sexe del pingüi.

Observació

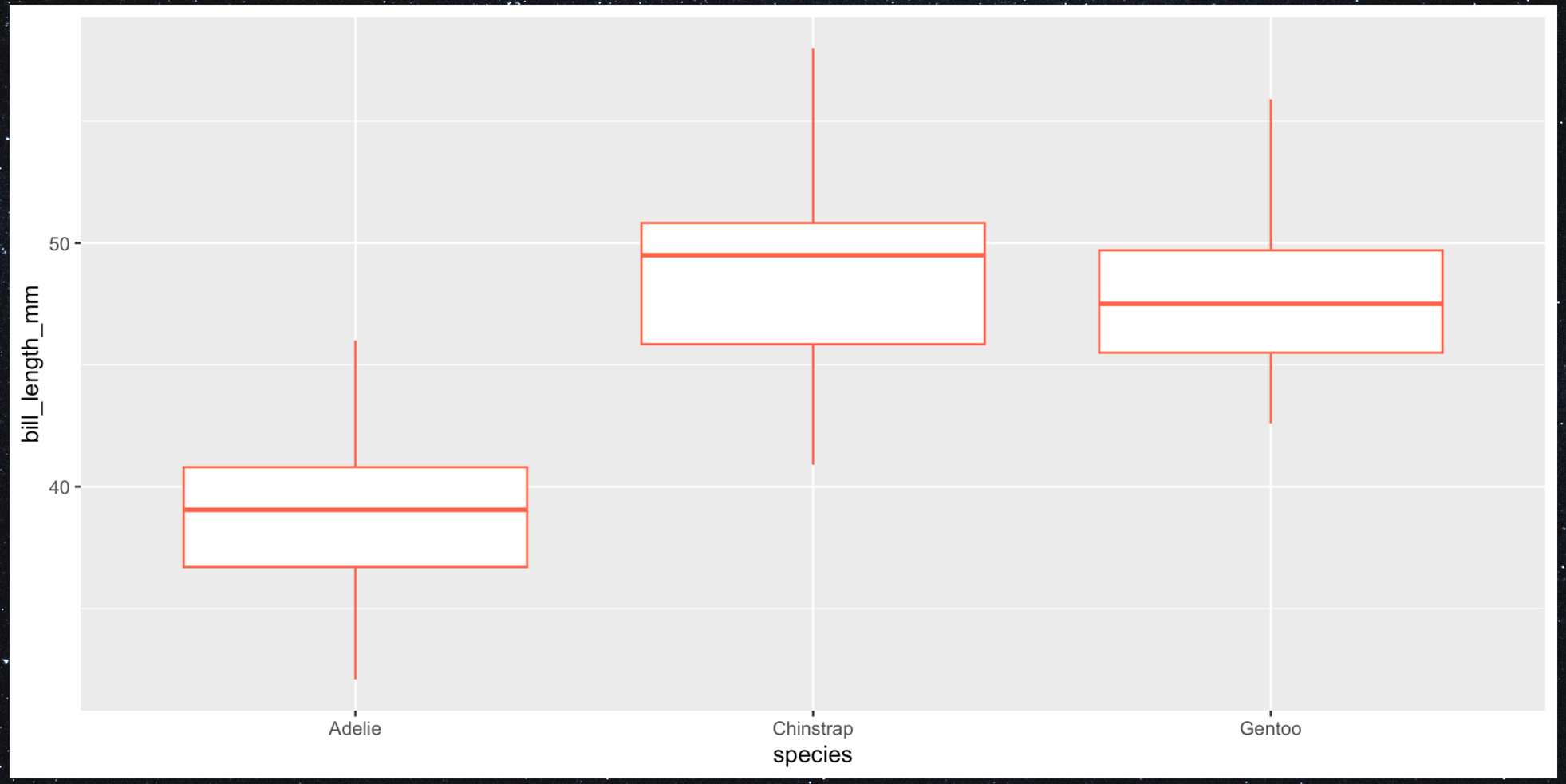
- Mentre que abans **volíem predir una variable numèrica** i contínua (el preu d'habitatge), ara **volem predir una variable categòrica**. Així doncs, haurem de fer servir un altre model diferent. En aquest cas farem servir un **model lineal generalitzat**.

preu
181000
124000
72000
137000
124500
228500
235000
204000
146300
350000

species
Gentoo
Gentoo
Adelie
Adelie
Adelie
Chinstrap
Gentoo
Gentoo
Adelie
Gentoo

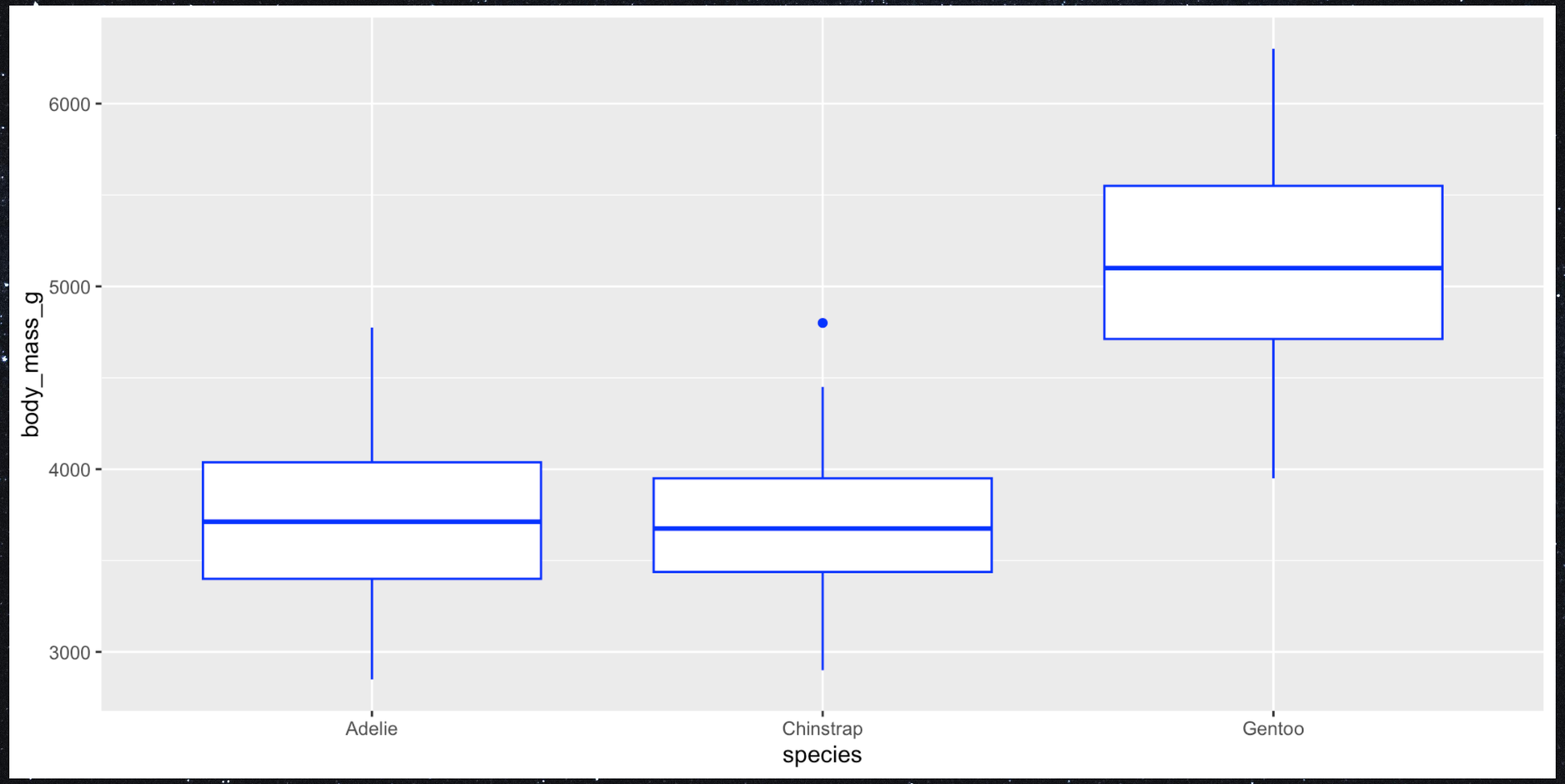
Exploració de dades

Boxplot espècies VS longitud de bec



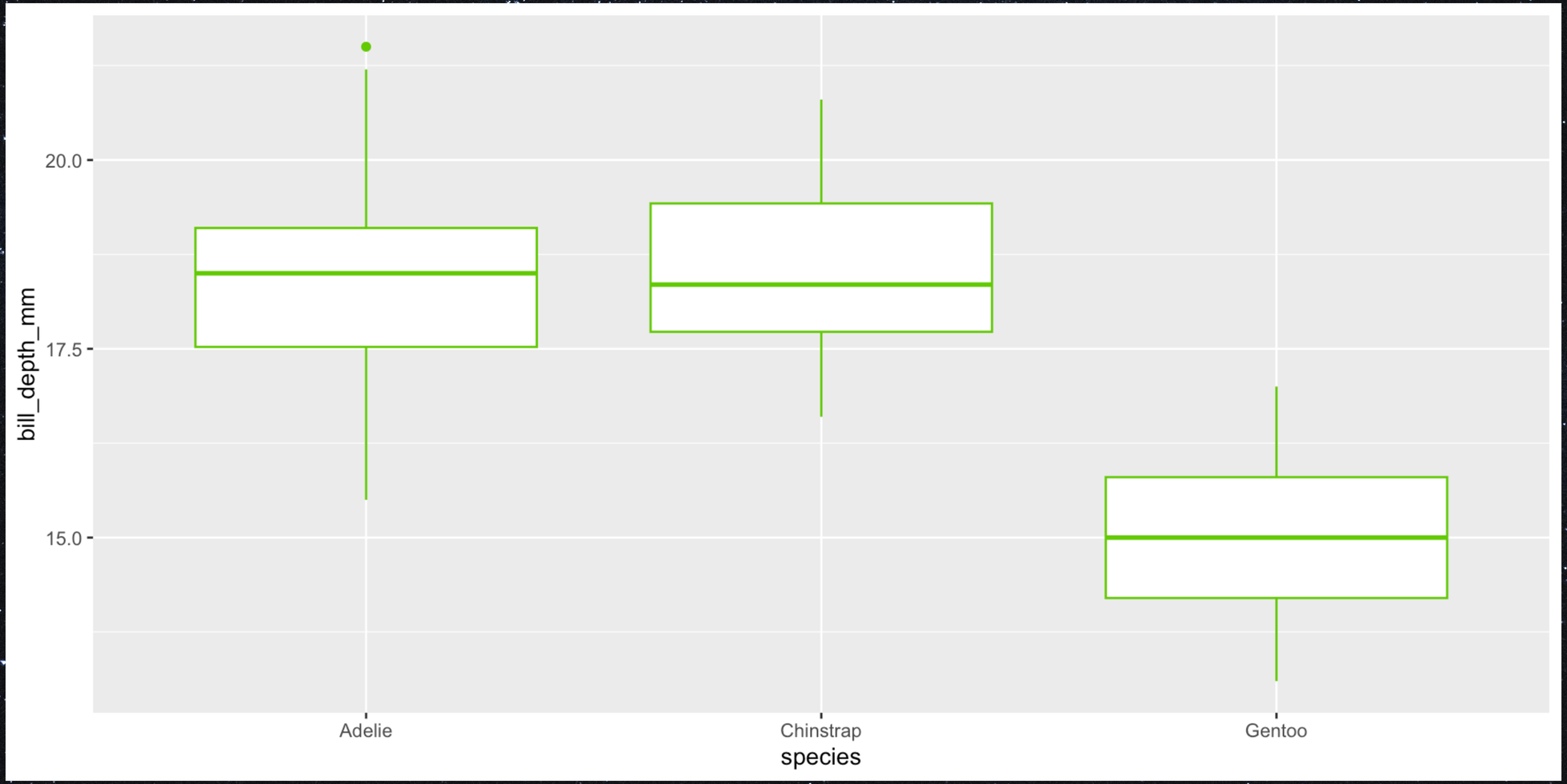
Exploració de dades

Boxplot espècies VS pes



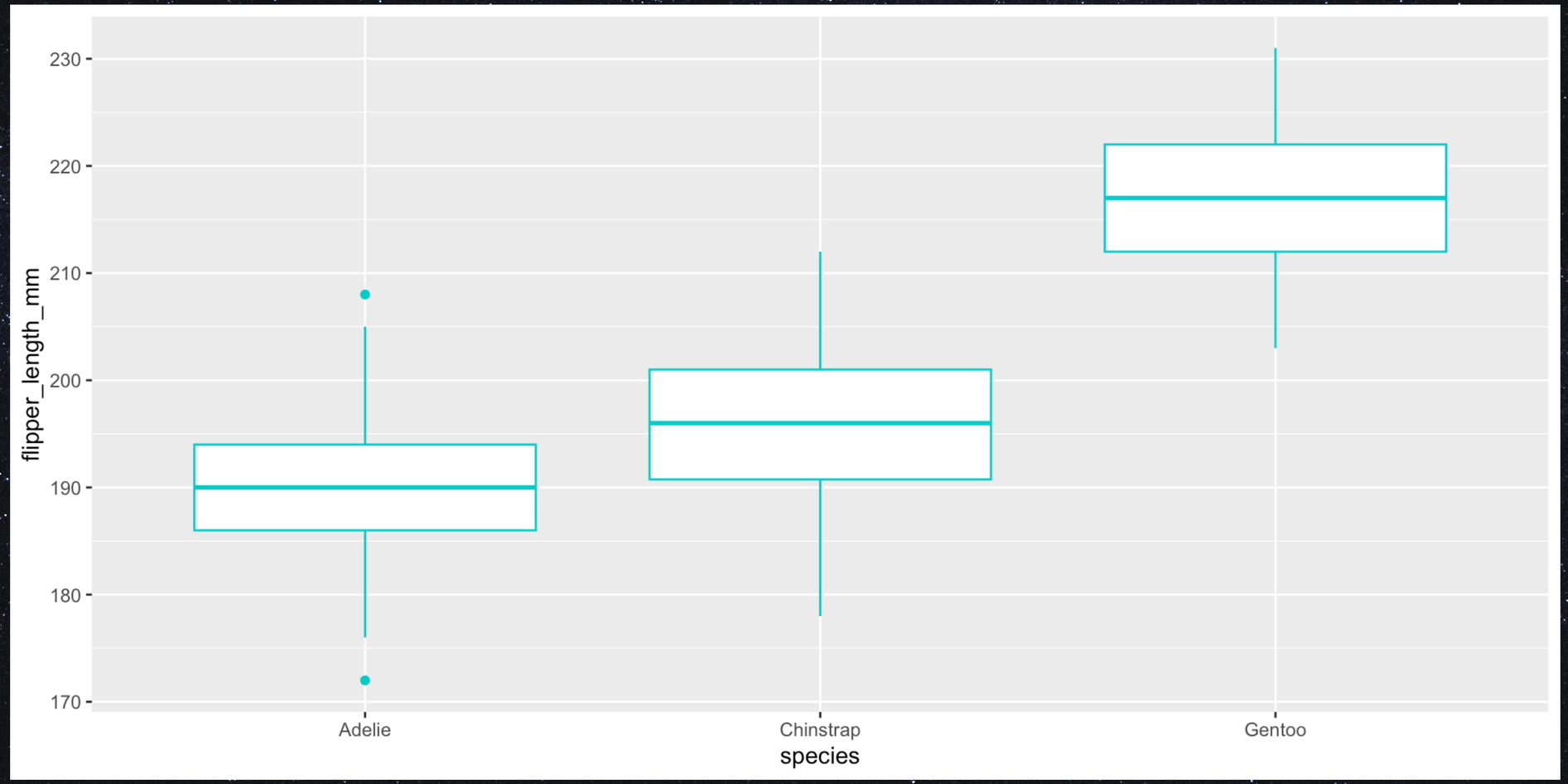
Exploració de dades

Boxplot espècies VS profunditat de bec



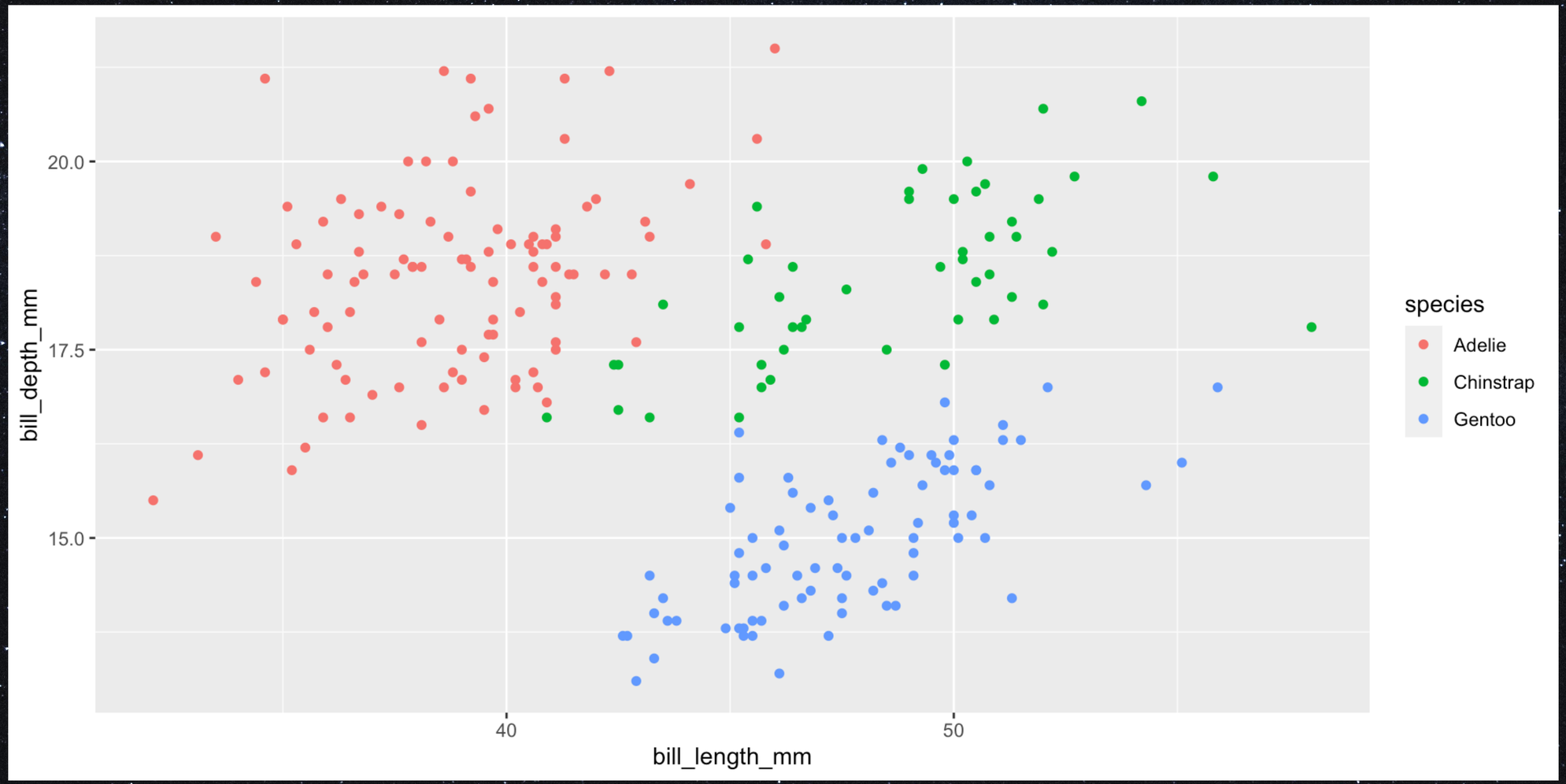
Exploració de dades

Boxplot espècies VS longitud de les ales



Exploració de dades

Gràfic de punts llargada de bec VS profunditat de bec



Model lineal generalitzat

Formulació i codi

The multinomial logistic regression assumes

$$y_i \mid x_i \overset{\text{ind.}}{\sim} \text{Multinomial}(1, (p_1(x_i), p_2(x_i), \ldots, p_k(x_i))), \qquad \sum_{j=1}^{k} p_j(x_i) = 1,$$

where

$$\log\left(\frac{p_j(x_i)}{p_1(x_i)}\right) = x_i' \beta_j, \qquad \beta_j \in \mathbb{R}^p, \; j \in \{2, \ldots, k\}.$$

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

Model lineal generalitzat

El model

- Ja tenim el nostre model ajustat, les variables usades i els coeficients estimats són:

```
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
```

Dades de test

- Fem servir el nostre model en la nova base de dades que els nostres companys ens han enviat. En aquesta taula de dades tenim 100 observacions.

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	38.9	17.8	181	3625	female	2007
Adelie	Torgersen	36.6	17.8	185	3700	female	2007
Adelie	Torgersen	42.5	20.7	197	4500	male	2007
Adelie	Biscoe	37.8	18.3	174	3400	female	2007
Adelie	Biscoe	38.2	18.1	185	3950	male	2007
Adelie	Biscoe	40.5	17.9	187	3200	female	2007
Adelie	Dream	37.2	18.1	178	3900	male	2007
Adelie	Dream	39.5	17.8	188	3300	female	2007
Adelie	Dream	36.4	17	195	3325	female	2007
Adelie	Dream	36	17.9	190	3450	female	2007

Prediccions

- A continuació, fem **prediccions** amb el model que hem entrenat anteriorment, i veiem si aquestes prediccions s'acosten al valor real de les noves dades.
- El codi per poder fer prediccions a R:

```
preds_train <- predict(model.lingen, newdata = penguins_test)
```

Resultats

- Alguns dels resultats de les prediccions del nostre model:

Real	Predit
Adelie	Adelie
Adelie	Adelie
Adelie	Adelie
Gentoo	Gentoo
Gentoo	Gentoo
Gentoo	Gentoo
Gentoo	Gentoo
Chinstrap	Chinstrap

Resultats

Matriu de confusió

- A la següent taula tenim les prediccions que ha fet el nostre model usant les noves dades.

Real

Predit

	Adelie	Chinstrap	Gentoo
Adelie	44	1	0
Chinstrap	0	19	0
Gentoo	0	0	36

Resultats

- 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.
- Podem veure que el nostre model ha fet una classificació amb una precisió del **99%**!

Accés lliure al repositori

- Podeu accedir lliurement a un repositori on hi ha penjades les dades i el codi de programació usat per fer-ho tot.

Exemples

- A la web podeu trobar molts exemples de models predictius.
  - Calculadora de risc cardiovascular: https://static.heart.org/riskcalc/app/index.html#!/baseline-risk
  - Calculadora de risc de càncer de mama: https://bcrisktool.cancer.gov/
  - Calculadora d'estalvis per a la jubilació: https://investor.vanguard.com/tools-calculators/retirement-income-calculator

Exemples

Models predictius en grans empreses i xarxes socials

- Google ofereix serveis d'anàlisi predictiu a empreses: https://cloud.google.com/learn/what-is-predictive-analytics?hl=es
- Empreses com Meta (Facebook) es mostren molt actives en recerca sobre Anàlisi de Dades https://research.facebook.com/publications/research-area/data-science/?s.

Exemples

Models predictius en grans empreses i xarxes socials

- **Recomanacions de contingut (Netflix, YouTube, Spotify):** Aquestes plataformes utilitzen models predictius per analitzar el comportament dels usuaris i oferir recomanacions personalitzades.
- **Publicitat personalitzada (Google, Facebook, Instagram):** Les empreses de xarxes socials utilitzen models predictius per mostrar anuncis rellevants per a cada usuari.
- **Anàlisi del sentiment i detecció de tendències (Twitter, TikTok):** Les xarxes socials utilitzen models predictius per analitzar en temps real els comentaris, tuits o publicacions dels usuaris, i detectar tendències o identificar l'opinió general (positiva, negativa o neutra) sobre un tema.