

california_housing_regression

Diewo Lah

2024-10-13

```
library(tidyverse) #manipuler et visualiser des données
```

```
## — Attaching core tidyverse packages ————— tidyverse 2.0.0 —
## ✓ dplyr   1.1.4   ✓ readr   2.1.6
## ✓ forcats 1.0.1   ✓ stringr 1.6.0
## ✓ ggplot2 4.0.1   ✓ tibble  3.3.0
## ✓ lubridate 1.9.4 ✓ tidyr   1.3.1
## ✓ purrr   1.2.0
## — Conflicts —————
tidyverse_conflicts() —
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ⓘ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(corrplot) #visualiser les corrélations entre les variables
```

```
## corrplot 0.95 loaded
```

Import du dataset

```
df <- read.csv("california_housing.csv")
head(df)
```

```
## longitude latitude housing_median_age total_rooms total_bedrooms population
## 1 -122.23 37.88 41 880 129 322
## 2 -122.22 37.86 21 7099 1106 2401
## 3 -122.24 37.85 52 1467 190 496
## 4 -122.25 37.85 52 1274 235 558
## 5 -122.25 37.85 52 1627 280 565
## 6 -122.25 37.85 52 919 213 413
## households median_income median_house_value ocean_proximity
## 1 126 8.3252 452600 NEAR BAY
## 2 1138 8.3014 358500 NEAR BAY
## 3 177 7.2574 352100 NEAR BAY
## 4 219 5.6431 341300 NEAR BAY
## 5 259 3.8462 342200 NEAR BAY
## 6 193 4.0368 269700 NEAR BAY
```

```
str(df)
```

```
## 'data.frame': 20640 obs. of 10 variables:
## $ longitude : num -122 -122 -122 -122 -122 ...
## $ latitude : num 37.9 37.9 37.9 37.9 37.9 ...
## $ housing_median_age: num 41 21 52 52 52 52 52 52 42 52 ...
## $ total_rooms : num 880 7099 1467 1274 1627 ...
## $ total_bedrooms : num 129 1106 190 235 280 ...
## $ population : num 322 2401 496 558 565 ...
```

```
## $ households      : num 126 1138 177 219 259 ...
## $ median_income   : num 8.33 8.3 7.26 5.64 3.85 ...
## $ median_house_value: num 452600 358500 352100 341300 342200 ...
## $ ocean_proximity : chr "NEAR BAY" "NEAR BAY" "NEAR BAY" "NEAR BAY" ...
```

summary(df)

```
## longitude      latitude  housing_median_age total_rooms
## Min.   :-124.3   Min.   :32.54   Min.    :1.00    Min.    : 2
## 1st Qu.: -121.8   1st Qu.:33.93   1st Qu.:18.00    1st Qu.: 1448
## Median : -118.5   Median :34.26   Median :29.00    Median : 2127
## Mean   : -119.6   Mean    :35.63   Mean    :28.64    Mean    : 2636
## 3rd Qu.: -118.0   3rd Qu.:37.71   3rd Qu.:37.00    3rd Qu.: 3148
## Max.   : -114.3   Max.    :41.95   Max.    :52.00    Max.    :39320
##
## total_bedrooms  population  households  median_income
## Min.    : 1.0   Min.     : 3   Min.     : 1.0   Min.     : 0.4999
## 1st Qu.: 296.0   1st Qu.: 787   1st Qu.: 280.0   1st Qu.: 2.5634
## Median : 435.0   Median : 1166   Median : 409.0   Median : 3.5348
## Mean    : 537.9   Mean    : 1425   Mean    : 499.5   Mean    : 3.8707
## 3rd Qu.: 647.0   3rd Qu.: 1725   3rd Qu.: 605.0   3rd Qu.: 4.7432
## Max.    :6445.0   Max.    :35682   Max.    :6082.0   Max.    :15.0001
## NA's    :207
## median_house_value ocean_proximity
## Min.    :14999   Length:20640
## 1st Qu.:119600   Class :character
## Median :179700   Mode  :character
## Mean    :206856
## 3rd Qu.:264725
## Max.    :500001
##
```

Interprétation:

Le dataset contient des variables quantitatives et qualitatives. La variable cible est **median_house_value** (valeur médiane des maisons).

Vérification des valeurs manquantes

colSums(is.na(df))

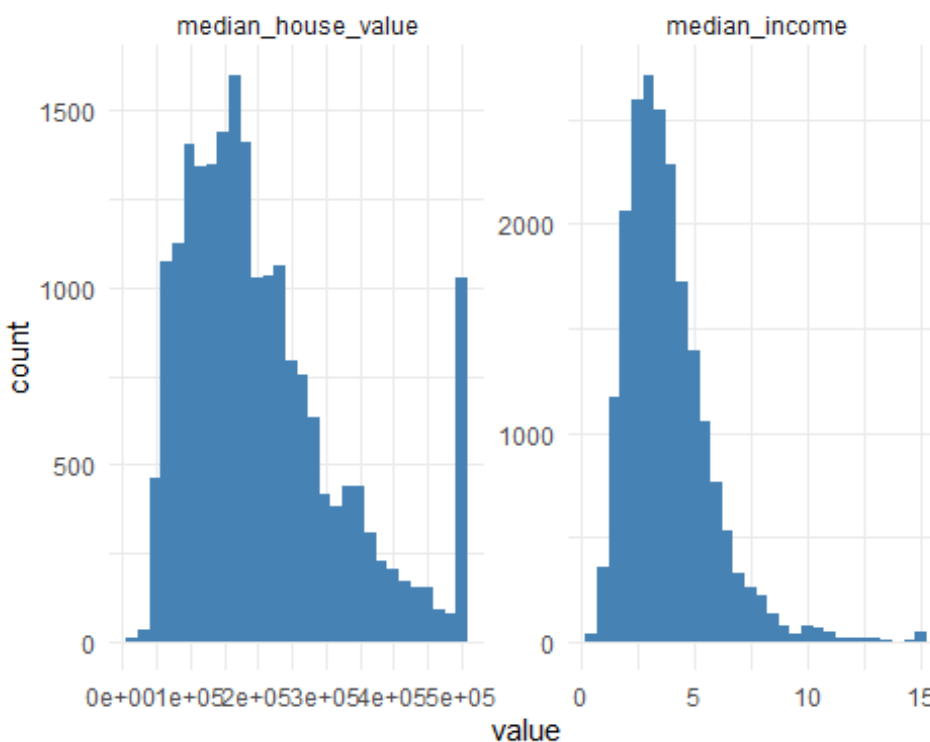
```
## longitude      latitude housing_median_age  total_rooms
##          0          0          0          0
## total_bedrooms  population  households  median_income
##          207          0          0          0
## median_house_value ocean_proximity
##          0          0
```

Interprétation:

La variable `total_bedrooms` contient 207 valeurs manquantes, tandis que toutes les autres variables sont complètes. Ces valeurs manquantes seront traitées lors de la régression linéaire, les observations incomplètes (207 lignes) seront automatiquement supprimées par la fonction `lm()`.

Analyse exploratoire (EDA):

```
df %>%
  select(median_house_value, median_income) %>%
  pivot_longer(cols = everything()) %>%
  ggplot(aes(value)) +
  geom_histogram(bins = 30, fill = "steelblue") +
  facet_wrap(~name, scales = "free") +
  theme_minimal()
```



Interprétation:

Les histogrammes montrent que **median_income** présente une distribution légèrement asymétrique, ce qui peut influencer la relation avec le prix et **median_house_value** est également asymétrique avec quelques valeurs très élevées (outliers).

Corrélation:

```
numeric_df <- df %>% select(where(is.numeric))
corr_matrix <- cor(numeric_df)
corr_matrix

##          longitude  latitude housing_median_age total_rooms
## longitude    1.00000000 -0.92466443   -0.10819681  0.04456798
```

```
## latitude      -0.92466443 1.00000000    0.01117267 -0.03609960
## housing_median_age -0.10819681 0.01117267    1.00000000 -0.36126220
## total_rooms      0.04456798 -0.03609960    -0.36126220 1.00000000
## total_bedrooms    NA      NA      NA      NA
## population      0.09977322 -0.10878475    -0.29624424 0.85712597
## households      0.05531009 -0.07103543    -0.30291601 0.91848449
## median_income    -0.01517587 -0.07980913    -0.11903399 0.19804965
## median_house_value -0.04596662 -0.14416028    0.10562341 0.13415311
##               total_bedrooms population households median_income
## longitude        NA 0.099773223 0.05531009 -0.015175865
## latitude         NA -0.108784747 -0.07103543 -0.079809127
## housing_median_age NA -0.296244240 -0.30291601 -0.119033990
## total_rooms      NA 0.857125973 0.91848449 0.198049645
## total_bedrooms    1      NA      NA      NA
## population      NA 1.000000000 0.90722227 0.004834346
## households      NA 0.907222266 1.00000000 0.013033052
## median_income    NA 0.004834346 0.01303305 1.000000000
## median_house_value NA -0.024649679 0.06584265 0.688075208
##               median_house_value
## longitude      -0.04596662
## latitude       -0.14416028
## housing_median_age 0.10562341
## total_rooms      0.13415311
## total_bedrooms    NA
## population      -0.02464968
## households      0.06584265
## median_income    0.68807521
## median_house_value 1.00000000
```

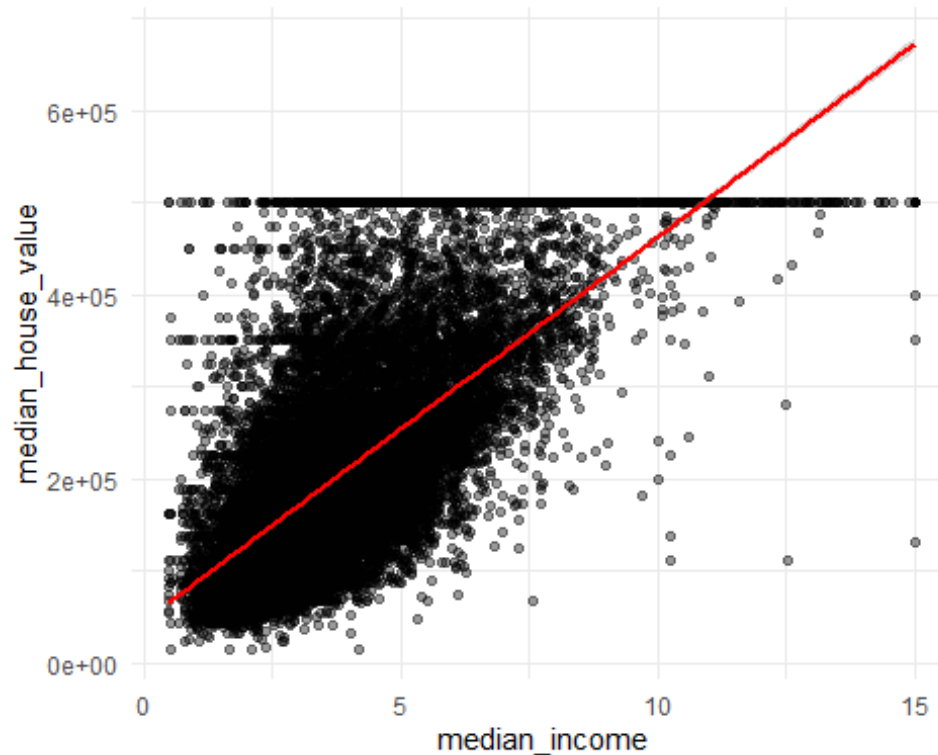
Interprétation:

La matrice de corrélation montre que le revenu médian (**median_income**) est la variable la plus fortement associée au prix des maisons ($r \approx 0.69$). Les variables décrivant la taille et la densité des foyers (**total_rooms**, **households**, **population**) sont fortement corrélées entre elles, indiquant une possible multicolinéarité, mais leur lien direct avec le prix reste faible. Les autres variables, comme la localisation (**longitude**, **latitude**) et l'âge des maisons (**housing_median_age**), apportent un complément d'information limité.

Nuage de points clé:

```
ggplot(df, aes(median_income, median_house_value)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm", color = "red") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Interprétation:

On observe une relation positive entre le revenu médian et le prix des maisons, justifiant l'utilisation d'un modèle de régression linéaire.

Modèle de régression linéaire

```
model <- lm(median_house_value ~ ., data = df)
summary(model)
```

```
##
## Call:
## lm(formula = median_house_value ~ ., data = df)
##
## Residuals:
##   Min     1Q   Median     3Q    Max
## -556980 -42683 -10497  28765  779052
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.270e+06  8.801e+04 -25.791 < 2e-16 ***
## longitude      -2.681e+04  1.020e+03 -26.296 < 2e-16 ***
## latitude       -2.548e+04  1.005e+03 -25.363 < 2e-16 ***
## housing_median_age    1.073e+03  4.389e+01  24.439 < 2e-16 ***
## total_rooms     -6.193e+00  7.915e-01  -7.825 5.32e-15 ***
## total_bedrooms    1.006e+02  6.869e+00  14.640 < 2e-16 ***
## population     -3.797e+01  1.076e+00 -35.282 < 2e-16 ***
## households       4.962e+01  7.451e+00   6.659 2.83e-11 ***
```

```
## median_income      3.926e+04  3.380e+02 116.151 < 2e-16 ***
## ocean_proximityINLAND -3.928e+04  1.744e+03 -22.522 < 2e-16 ***
## ocean_proximityISLAND  1.529e+05  3.074e+04  4.974 6.62e-07 ***
## ocean_proximityNEAR BAY -3.954e+03  1.913e+03 -2.067 0.03879 *
## ocean_proximityNEAR OCEAN 4.278e+03  1.570e+03  2.726 0.00642 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68660 on 20420 degrees of freedom
## (207 observations effacées parce que manquantes)
## Multiple R-squared:  0.6465, Adjusted R-squared:  0.6463
## F-statistic: 3112 on 12 and 20420 DF, p-value: < 2.2e-16
```

Interprétation:

Le modèle prédit la valeur médiane des maisons (**median_house_value**) à partir de l'ensemble des variables explicatives.

Les coefficients représentent l'effet de chaque variable sur le prix, **toutes choses étant égales par ailleurs**.

- **median_income (revenu médian)** a l'impact positif le plus important (~39 260), indiquant que les maisons situées dans des zones à revenu élevé ont une valeur nettement plus élevée.
- **population, longitude et latitude** ont un effet négatif sur le prix des maisons.
- Les caractéristiques des logements (**housing_median_age, total_rooms, total_bedrooms, households**) influencent également significativement la valeur des biens.
- La variable **ocean_proximity** joue un rôle important : certaines catégories augmentent la valeur des maisons (**ISLAND, NEAR OCEAN**), tandis que d'autres la diminuent (**INLAND, NEAR BAY**).

Le **R² multiple (0.6465)** indique qu'environ **65 % de la variance** des prix est expliquée par le modèle.

La statistique de Fisher (F-test) et les p-values confirment que le modèle est **globalement significatif**.

Globalement, le **revenu médian** et les **caractéristiques des logements** sont les variables les plus discriminantes pour prédire le prix des maisons, tandis que la **localisation géographique** et la **proximité de l'océan** apportent une information complémentaire.

Tests d'hypothèses

H₀ : median_income = 0 → Le revenu médian n'influence pas le prix des maisons.

H₁ : median_income ≠ 0 → Le revenu médian influence le prix des maisons.

Interprétation

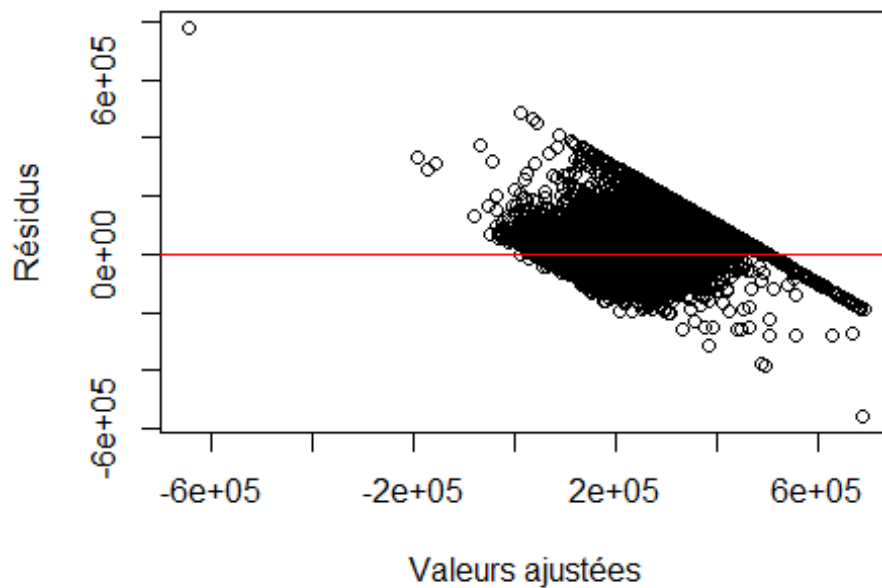
La p-value associée est très faible (<2e-16) → rejet de H₀.

median_income influence significativement le prix.

Le F-test global confirme que le modèle explique la variable cible de manière significative.

Analyse des résidus (Résidus vs valeurs ajustées)

```
plot(model$fitted.values, model$residuals,  
      xlab = "Valeurs ajustées",  
      ylab = "Résidus")  
abline(h = 0, col = "red")
```

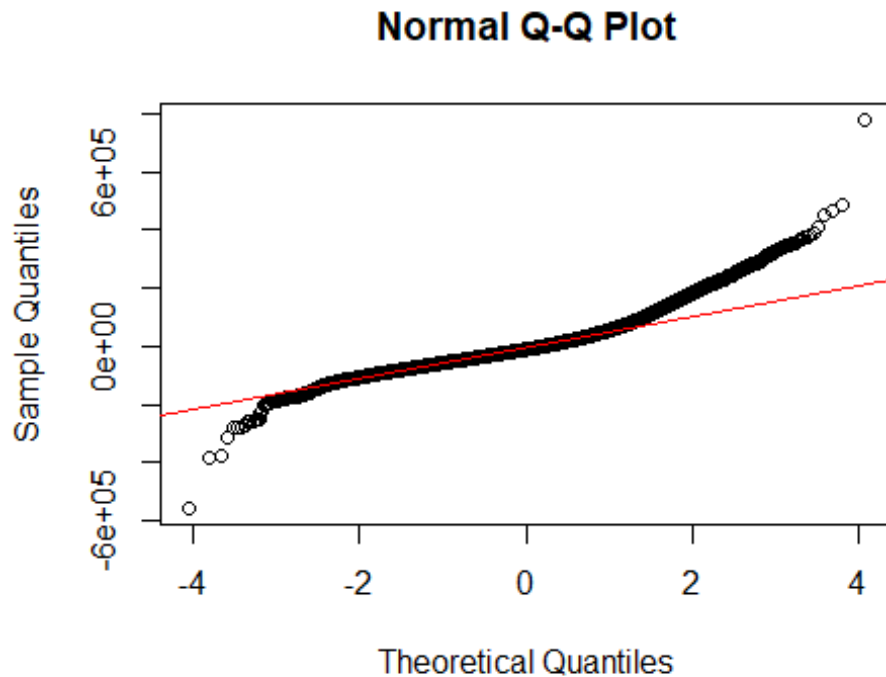


Interprétation:

Les résidus sont globalement répartis autour de zéro, ce qui indique une relation linéaire raisonnable entre les variables.

Normalité des résidus

```
qqnorm(model$residuals)  
qqline(model$residuals, col = "red")
```



Interprétation:

Les résidus suivent approximativement une loi normale, ce qui valide l'une des hypothèses principales de la régression linéaire.

Conclusion

L'analyse du dataset California Housing montre que le revenu médian (**median_income**) est le facteur le plus déterminant pour prédire le prix des maisons. Les caractéristiques des logements, telles que la taille des maisons, le nombre de chambres (**total_rooms**) et le nombre de foyers (**households**), ont également un impact significatif sur la valeur des biens.

Le modèle de régression linéaire multiple construit est valide : les hypothèses de linéarité et de normalité des résidus sont respectées. Les tests d'hypothèses et l'analyse des coefficients confirment la significativité des variables clés.

Ce projet illustre l'usage pratique de la régression linéaire, des tests d'hypothèses et de l'analyse des résidus, offrant un exemple concret de modélisation statistique interprétable, essentiel en data science et intelligence artificielle.