HOGESCHOOL UTRECHT

Fundamentals of Machine Learning
Week 3: linear regression

Jonas Moons

Check-in

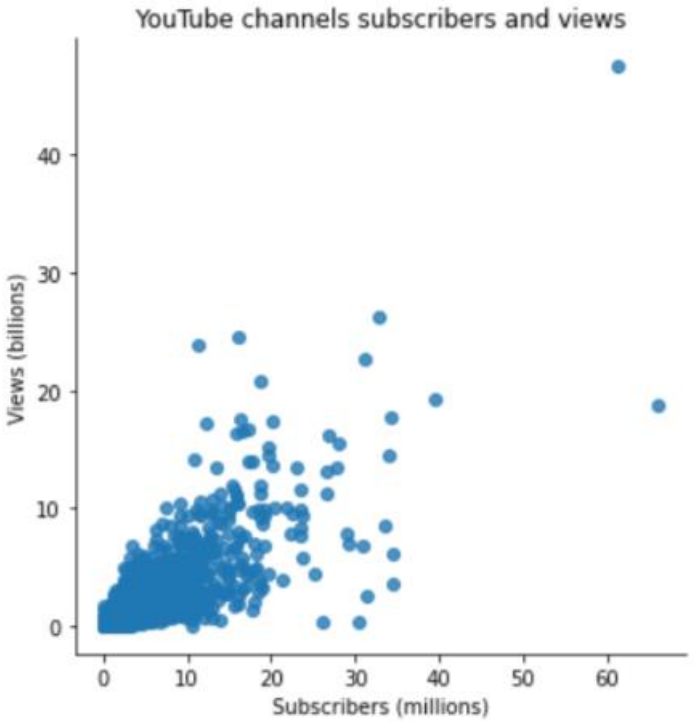
Feedback

- Use Markdown for (plenty of) comments, including headers: introduction, process, commenting on results
- Make sure to add titles and clear axes labels (not just the name of the variable in Pandas)

Topics

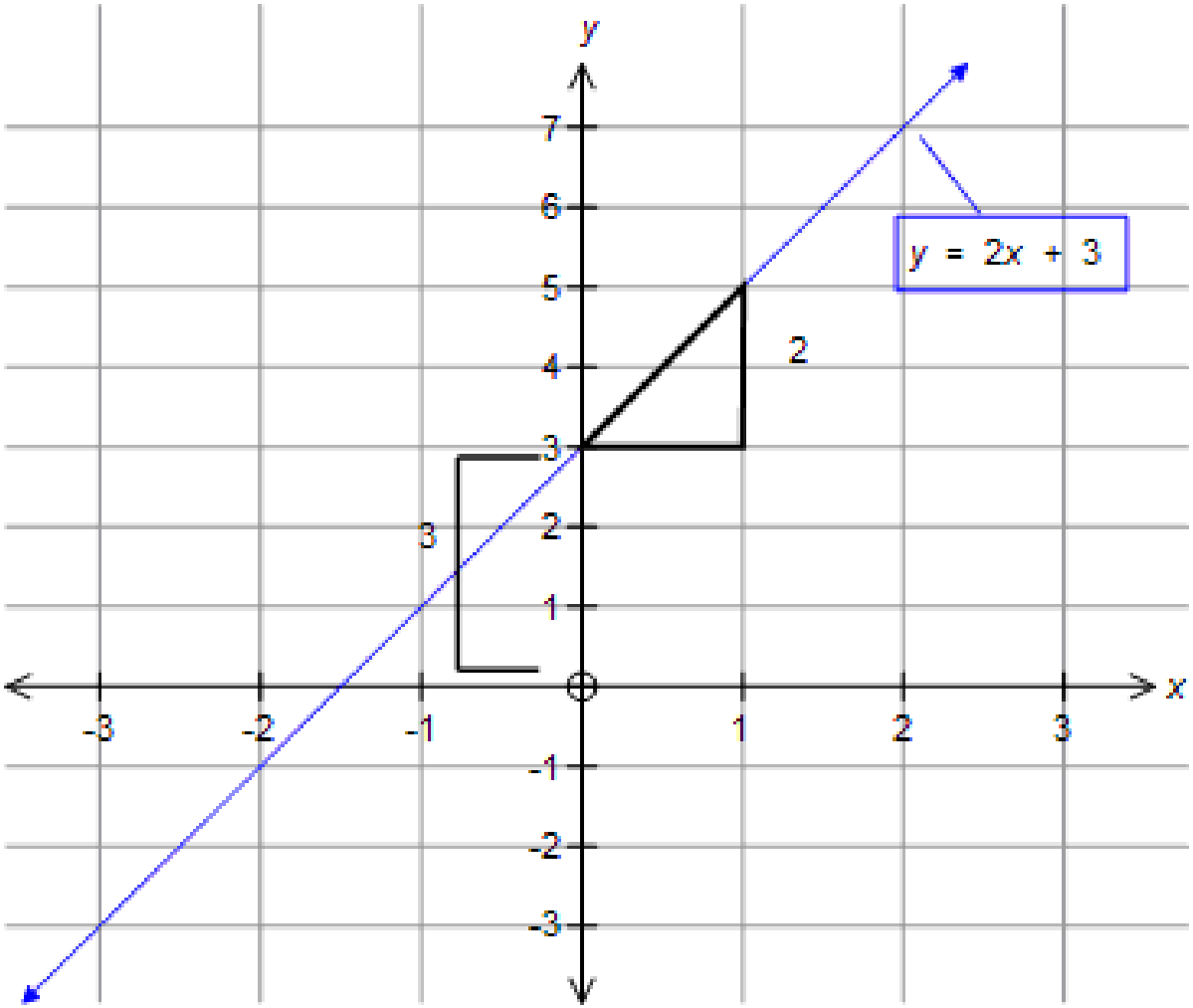
- Simple linear regression
- Evaluation
- Overfitting and model validation
- Multiple linear regression

Linear relation



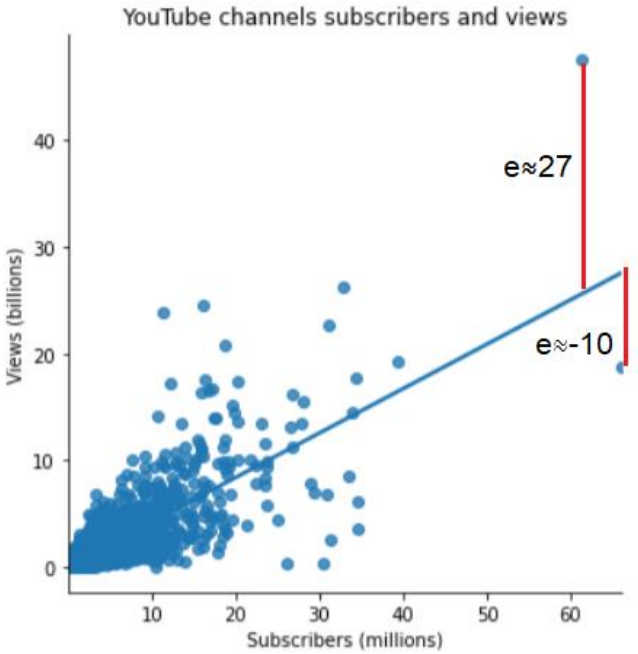
- The points in the scatter plot seem to be centered around a line, with some variation
- Variance in views seems to *increase* with subscribers

Linear relation



Linear regression formula

$$Y = b_0 + b_1 X + e$$



Y = dependent variable: views
X = independent variable: subscribers

Coefficients (constant):
b_0 = intercept: how many views with no subscribers = 0.046 billion
b_1 = slope: views per million subscribers = 0.417 billion
(=417 views per subscriber)

e = error / residual: what's left over, what we can't explain

The algorithm that finds the line **minimizes the *squared* errors**

Example data points

$$Y = \mathrm{b}_0 + \mathrm{b}_1 X + e$$

Channel	X (million subscribers)	Y (billion views)	b_0	b_1	Y' (predicted views)	e (residual)
PewDiePie	66	19	0.046	0.417	28	-9

(Numbers are rounded down here and there to make it clearer)

Example data points

$$Y = b_0 + b_1 X + e$$

Channel	X (million subscribers)	Y (billion views)	b_0	b_1	Y' (predicted views)	e (residual)
PewDiePie	66	19	0.046	0.416	28	-9
Taylor Swift	28	16	0.046	0.416	12	4

(Numbers are rounded down here and there to make it clearer)

Exercise 1: linear regression

See the *linear_regression* Notebook in the Examples folder.

Using Seaborn, Pandas and sk-learn:

1. Import the Funda data set (or continue in the same Notebook as from last week)

2. Clean up the *price* variable. Make a scatter plot of price and surface area, both with and without a regression line. Which should be X and which should be Y?

3. Train a linear model for the relation using sk-learn. Write down the complete formula with X, Y and the coefficients b_0 and b_1 (as numbers). Use a Markdown cell and write the formula in pretty math notation.

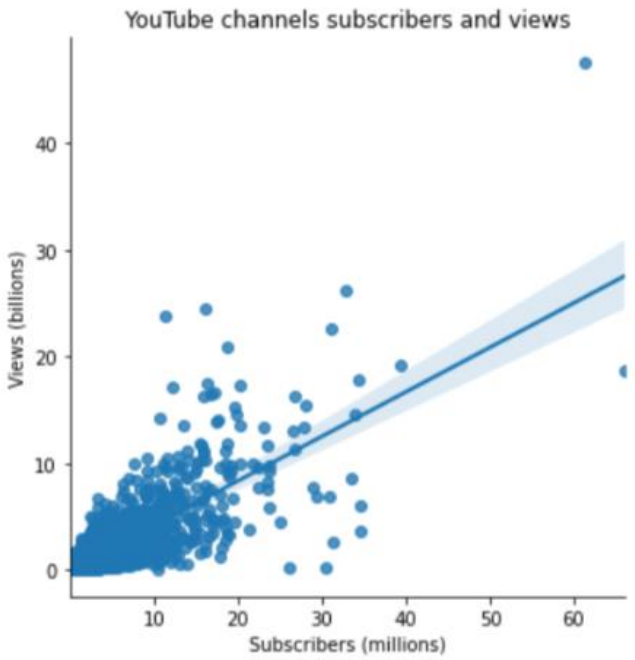
4. Predict the price for a house of 70 m^2

5. Use sk-learn to predict the price for all the houses in the dataset. What is the prediction for Slichtenhorststraat 10? What is the residual? Note: how to get this value is not in the example Notebook, but perhaps you can figure out how to retrieve it.

Topics

- Simple linear regression
- Evaluation
- Overfitting and model validation
- Multiple linear regression

Assumptions of simple linear regression



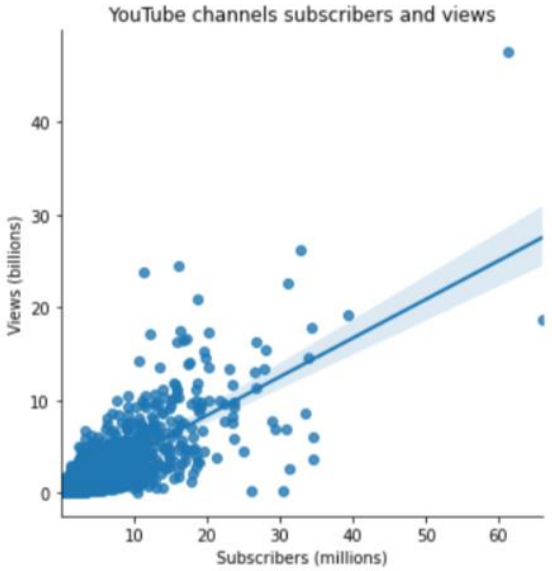
1. Linearity (the points are around a straight line, not on a curved line)

2. Equal variance (the distance between the points and the line does not change very much)

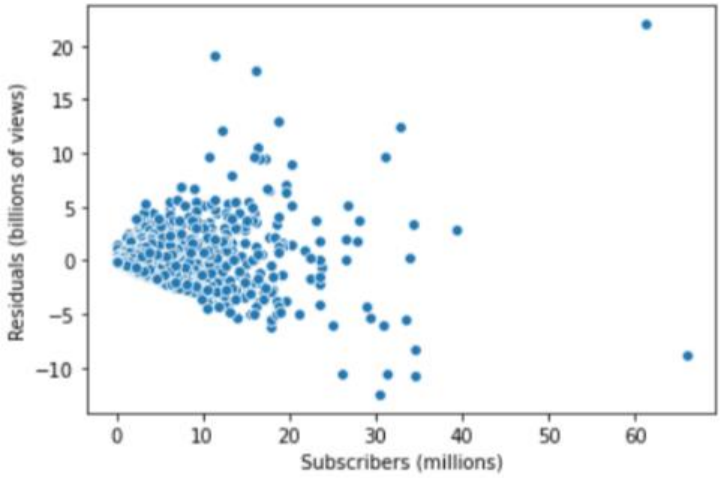
3. Residuals are normally distributed (ignore this one for now)

Inspect residuals/errors

Linear model



Residuals (distance from the line)

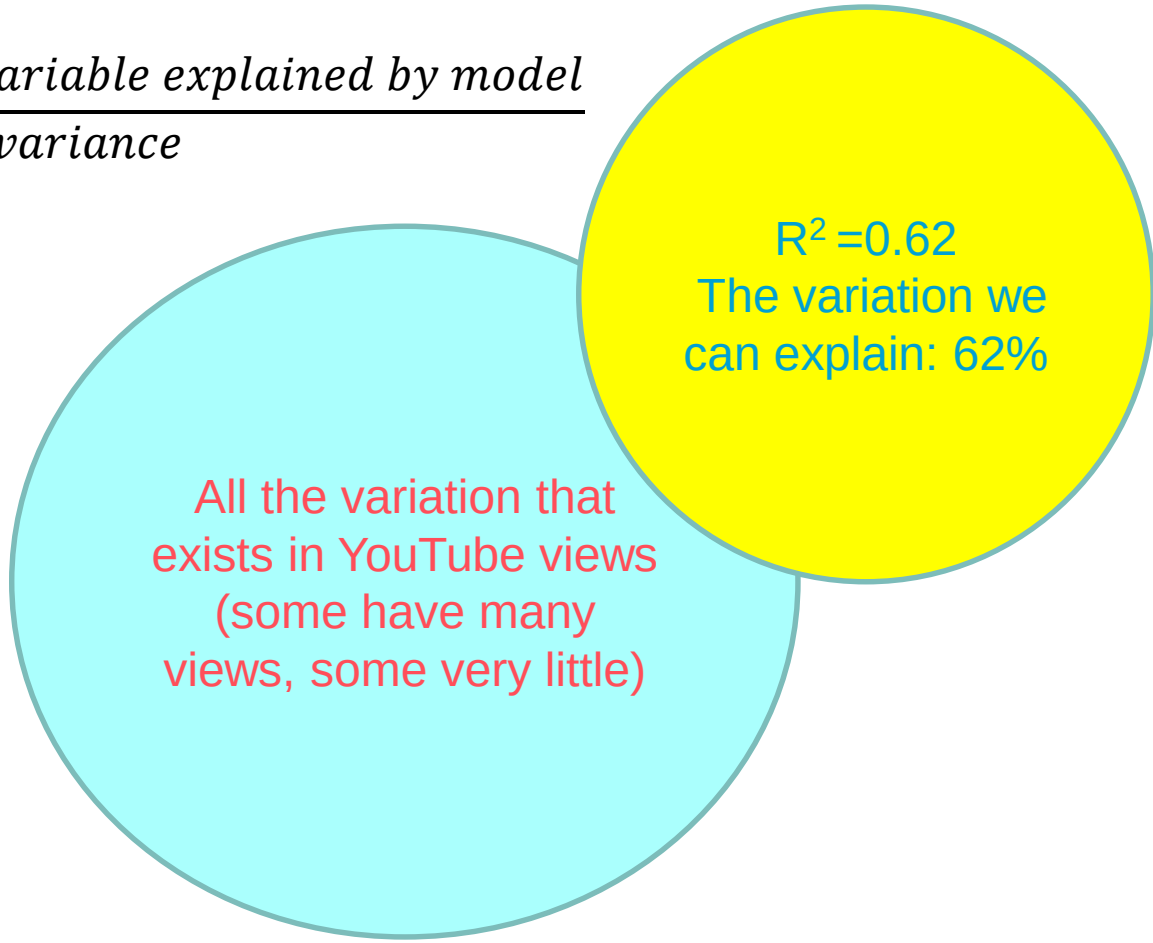


This means our predictions are poor for high numbers of subscribers!

Model fit: R^2

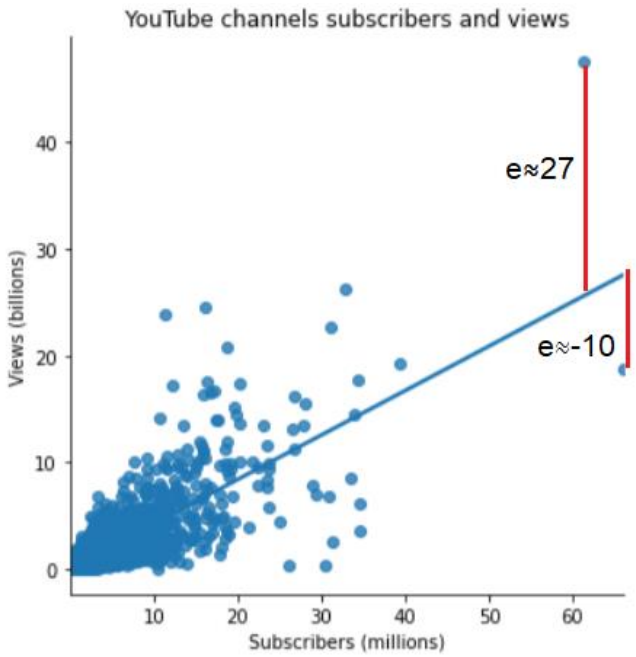
$$R^2 = \frac{variance\ in\ dependent\ variable\ explained\ by\ model}{total\ variance}$$

- Varies from 0 to 1
- The proportion of variance that you can explain with your model



Model fit: RMSE

- **R**oot **M**ean **S**quared **E**rror
- Take the mean of all the squared errors/residuals (all the ‘sticks’)
  - Then take the root (√) of that
- ≈ “How much is your prediction typically off”



Exercise 2: model evaluation

Continue in your Notebook. We are going to evaluate the fit of the model, with:

X: surface area in m^2

Y: price in euros

1. Calculate the residuals (e) and add them to the dataframe

2. Make the following plots:
   - A scatter plot of (Y, Y') (Y' means predicted Y)
   - A scatter plot of (X, e)

3. What is the R^2 of the model?

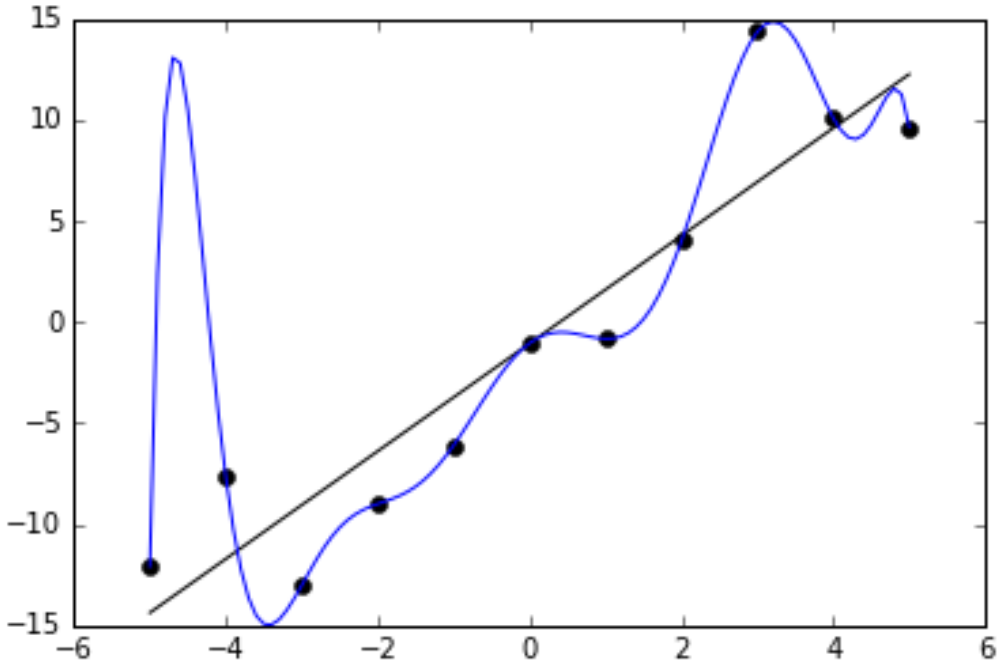
4. What is the root mean squared error (RMSE)?

5. What is your conclusion about the fit? Which houses are predicted better? Cheap or expensive house?

Topics

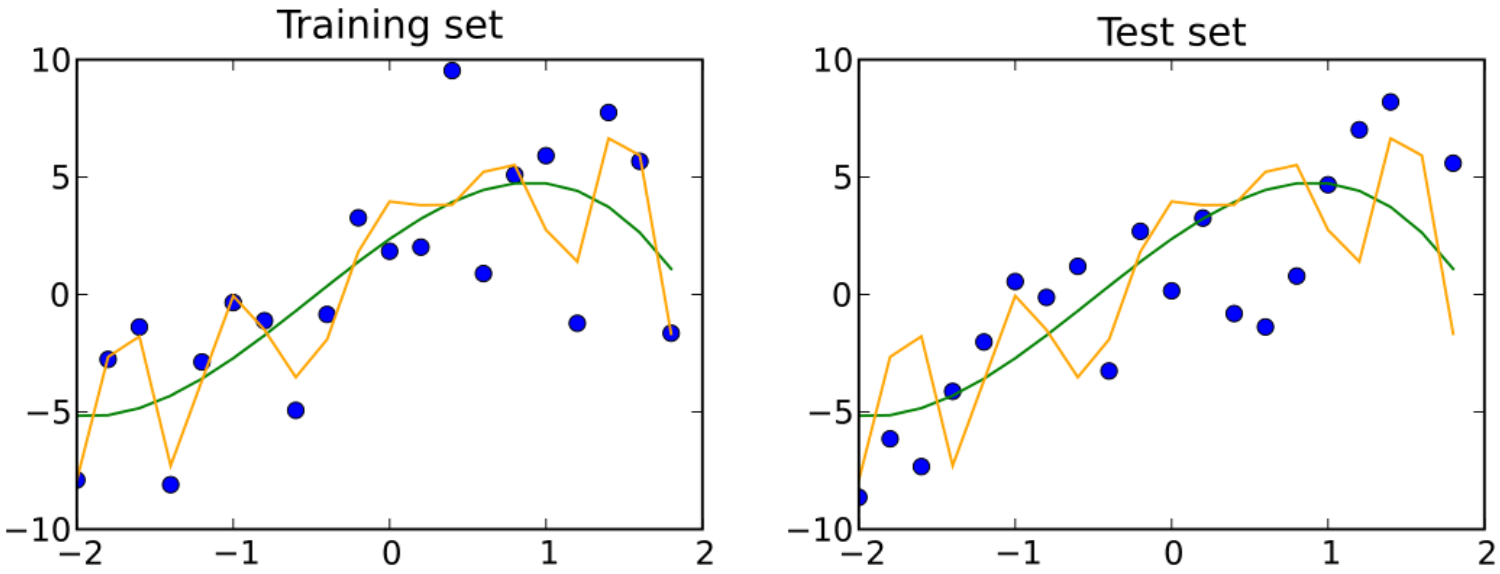
- Simple linear regression
- Evaluation
- Overfitting and model validation
- Multiple linear regression

Overfitting



Even though the curved line fits the sample data well, it’s probably a poor model for the population! The line fits the population much better.

Training and test set



- Two models trained on the training data set
  - Simple model in green
  - Complex model in orange

- The test set is used only to *test* the data (not to calculate the coefficients / train the model).
  - The orange model overfits: it fits the training set ‘too well’

Training and test set

- Usually in machine learning, we split the data at the beginning into a training (70-80%) and test set (20-30%)
- We train the model on the training set
- Then, *using the model coefficients from the training set,* we report performance **on the test set**
- So, in linear regression, we take the line we found with the train data and plot the test data around it

Topics

- Simple linear regression
- Evaluation
- Overfitting and model validation
- Multiple linear regression

Equation

$$Y = \mathrm{b}_0 + \mathrm{b}_1 X_1 + \mathrm{b}_2 X_2 + \cdots + e$$

For instance:

$$wage = 2 + 0.3 \cdot age + 0.2 \cdot years\_experience + \cdots + e$$

Categorical variables

- Linear regression and other algorithms only accept numerical variables. So how can we use categorical variables?

- By creating so-called ‘dummy variables’. Pandas can create them automically with pd.get_dummies()

title	action	comedy	drama
Guardians of the Galaxy	1	0	0
Prometheus	0	0	0
Split	0	0	0
Sing	0	1	0
Suicide Squad	1	0	0
The Great Wall	1	0	0

Exercise 3

For reference, use the example Notebook *multiple_linear_regression*

1. Make a scatterplot matrix and a correlation matrix of the *Funda* data set

2. Based on this, choose 3 variables from the Funda data set as your independent variables (X) for the dependent variable price (Y). You want variables that have a linear relation and correlate well with *price*. Note: it does not work very well to select variables that are highly correlated ***with each other***, such as *rooms* and *bedrooms*.

3. Split the data into a test set and a train set

4. Evaluate model performance by calculating R^2 and RMSE *on the test set*

5. Make a plot of the predicted price vs. the actual price. Which prices are predicted well?

6. Print out the coefficients and write the result in plain English.