**Regularization Methods: Ridge & Lasso**

* **Supervised Learning**

The project consisted of predicting sales based on ad spending.

**Import libraries**

Like with any project, we import our usual libraries that will help us perform basic data manipulation and plotting.



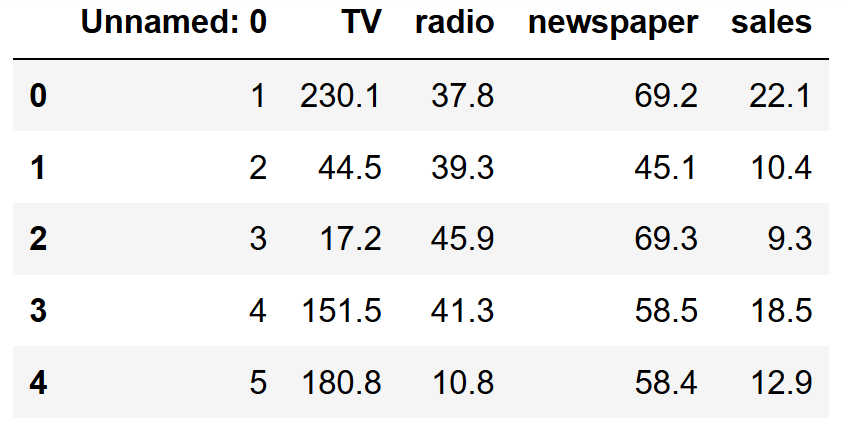
Now, we can start our exploratory data analysis.

**Exploratory data analysis**

We start off by importing our dataset and looking at the first five rows:



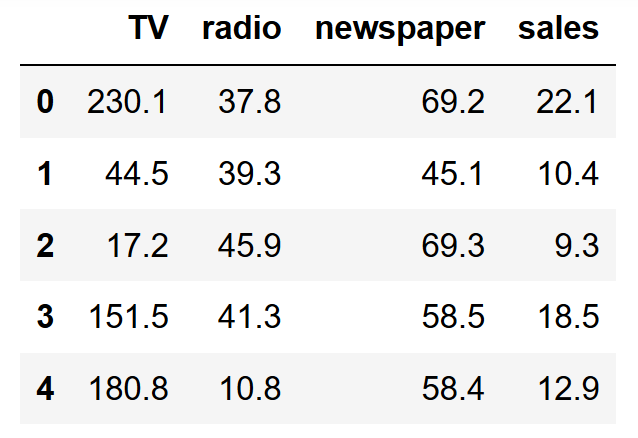
You should see:



Notice that the *Unnamed: 0*column is useless. Let’s take it out.



And now, our dataset looks like this:



As you can see, we only have three advertising mediums, and *sales*is our target variable.

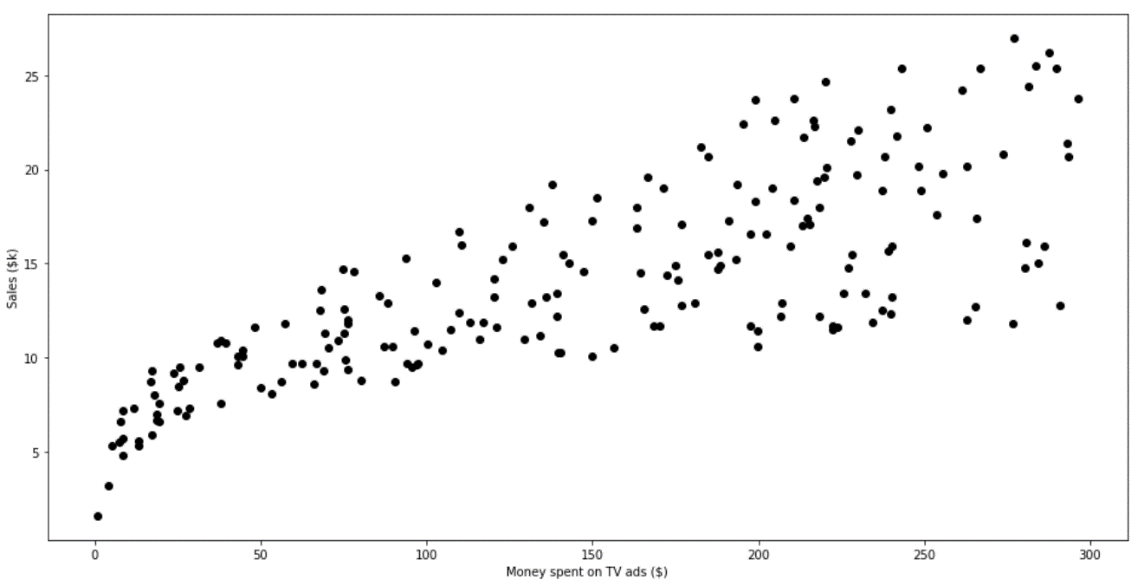
Let’s see how each variable impacts the sales by making a scatter plot. First, we build a helper function to make a scatter plot:



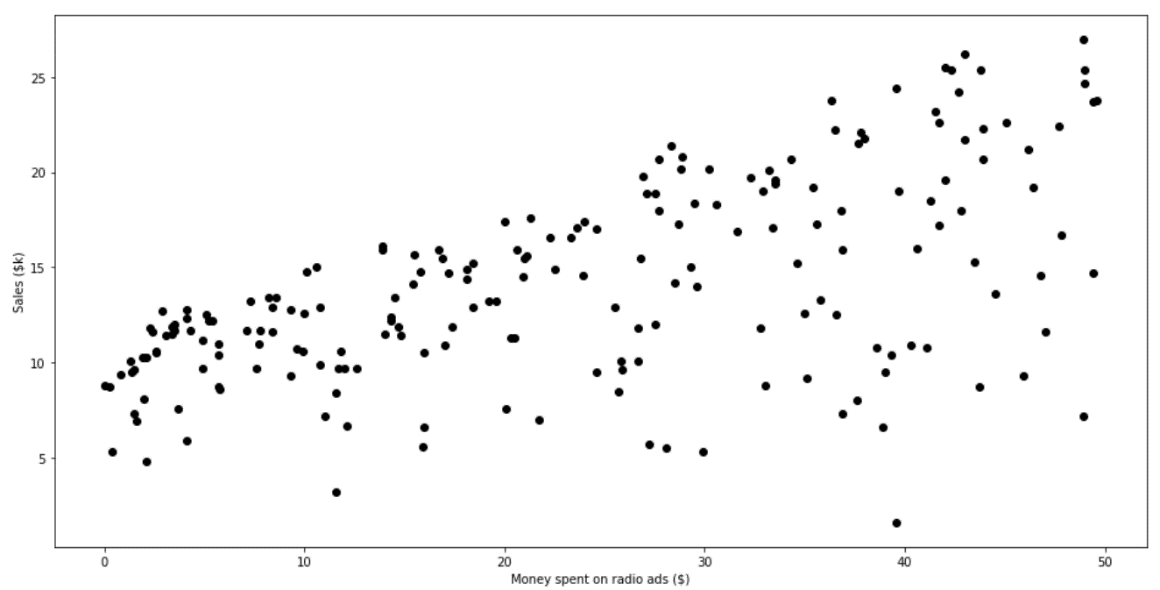
Now, we can generate three different plots for each feature.



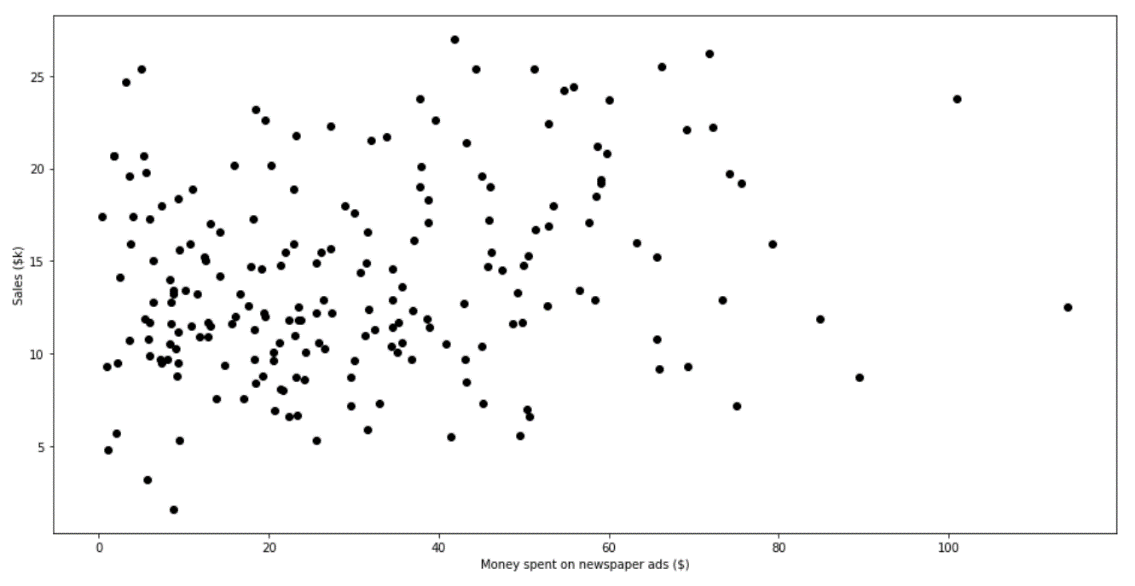
And you get the following:



Sales with respect to money spend on TV ads



Sales with respect to money spent on radio ads



Sales with respect to money spent on newspaper ads

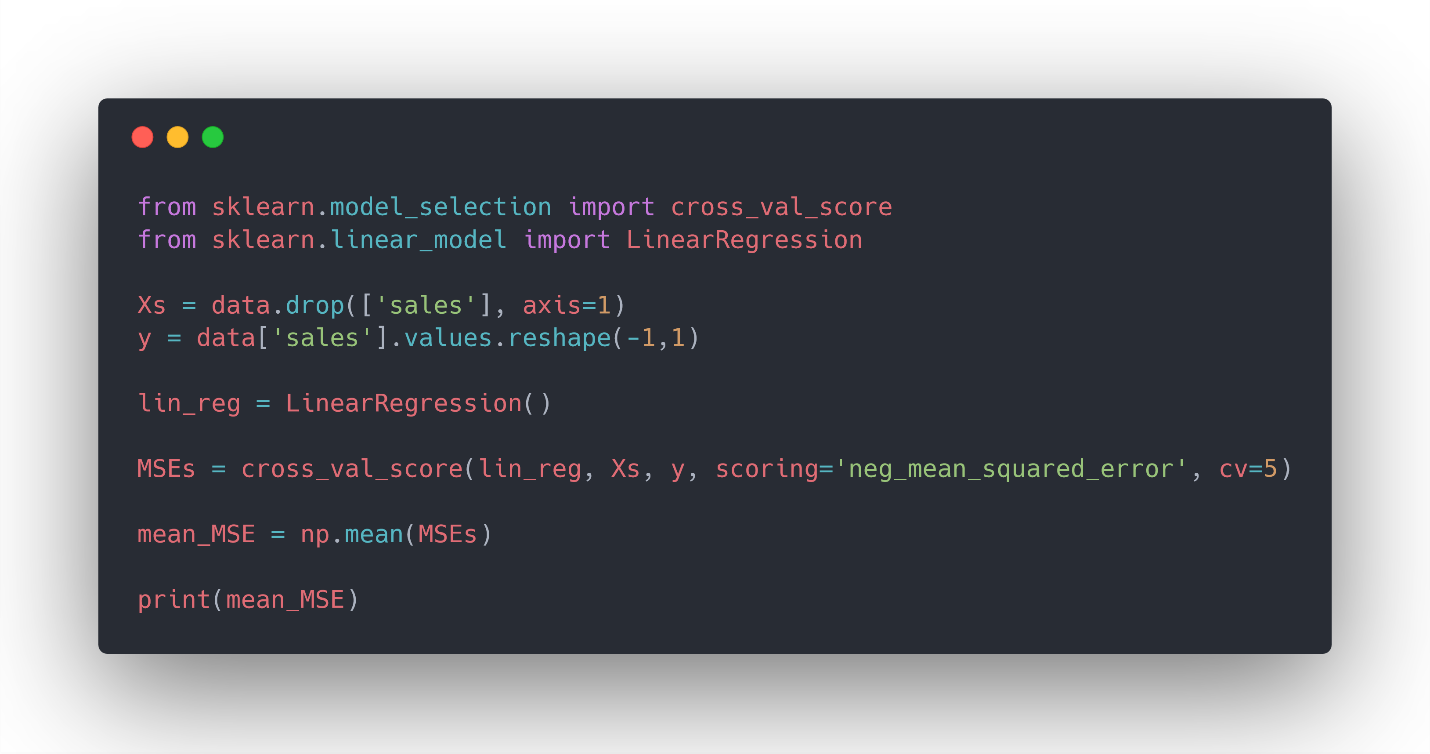
As you can see, TV and radio ads seem to be good predictors for sales, while there seems to be no correlations between sales and newspaper ads.

Luckily, our dataset does not require further processing, so we are ready to move on to modelling right away!

**Modelling**

**Multiple linear regression — least squares fitting**

Let’s take a look at what the code looks like, before going through it.



First, we import the *LinearRegression* and *cross\_val\_score*objects. The first one will allow us to fit a linear model, while the second object will perform k-fold cross-validation.

Then, we define our features and target variable.

The *cross\_val\_score*will return an array of MSE for each cross-validation steps. In our case, we have five of them. Therefore, we take the mean of MSE and print it. You should get a negative MSE of -3.0729.

Now, let’s see if ridge regression or lasso will be better.

**Ridge regression**

For ridge regression, we introduce *GridSearchCV*. This will allow us to automatically perform 5-fold cross-validation with a range of different regularization parameters in order to find the optimal value of *alpha*.

The code looks like this:



Then, we can find the best parameter and the best MSE with the following:



You should see that the optimal value of *alpha*is 20, with a negative MSE of -3.07267. This is a slight improvement upon the basic multiple linear regression.

**Lasso**

For lasso, we follow a very similar process to ridge regression:



In this case, the optimal value for *alpha* is 1, and the negative MSE is -3.0414, which is the best score of all three models!