# Exercise 6 - Data Cleaning with Pandas

Load in the dataset renfe\_trains.csv

### Initial Data Inspection

Inspect the columns of the DataFrame. Specifically, consider the type of each column and whether it seems reasonable. If not, imvestigate why.

It seems like we have some bad values in the price column with the value ‘price’. Let’s inspect a the rows where this is the case.

It looks like some sort of error has meant the column names have been fed into the data in intervals. Let’s drop these rows as they are clearly an accident.

We can now represent price as having the appropriate type. Convert it so that it does.

### Missing Values

Identify whether there are missing values in the DataFrame.

Which columns are they in?

Inspect some rows which contain them.

Drop all rows with missing values across all three columns vehicle\_class, price, and fare

Run the below code. What does it sugest about ticket price with respect to vehicle\_class and fare?

df[['vehicle\_class', 'fare', 'price']].groupby(['vehicle\_class', 'fare']).mean()

For now, we want you to fill the missing price with the mean of all prices. However, in the extention, you can try to tackle this more correctly (and trickily!).

Fill the missing price values with the mean of all prices.

Check you have gotten rid of all NaN values.

### Deduplication

Use duplicated to see whether the dataset contains any duplicated rows

As the dataset constitutes ticket price search results, theres a good chance duplication has come about due to the data collection method. For example, there are many tickets available on each train. We would want to investigate this further, but to use the functionality, lets get rid of these duplicate rows.

### Data Transformation

Now that we have a dataset without the above errors, we want to add some meaningful columns.

First, lets bin price into three groups and add that as column to the DataFrame called price\_bin.

The bins should be [(0, 150], (150, 300], (300, 450]].

Plot a bar chart of the binned prices

Compute the amount of tax that would need to be paid if vat is at 20% and prices aren’t inclusive of it.

### Text Data Manipulation

Capitalise the word renfe in the company column

Generate a statistical summary of trips whose vehicle\_class contains Turista

Display all rows whose fare ends with a +

### Stretch Exercises

As it appears price depends upon vehicle\_class and fare, we choose to replace missing price values with the average for their vehicle\_class and fare category. Write some code which does this.