# Case Study 1: Preparing and Analysing Data

In your lab environment, load in the dataset renfe\_trains.csv with an appropriate method and create a new notebook to work with the data, completing the following tasks.

### Initial data inspection and loading issues

1. Inspect the columns of the Data. Specifically, consider the detected data type of each column and whether it seems reasonable. If not, investigate why.
2. You may see some bad values in the price column with the value ‘price’.   
   Inspect the specific rows where this is the case.
3. It is possible a loading error means the column names have been fed into the data in intervals. Filter out any rows that have been loaded incorrectly.
4. Following this fix, the price column can now use the appropriate data type.

### Statistical analysis and observations.

1. Using any method, calculate mean and median, as well as standard deviation and inter quartile range of the price column.
2. Visualise the price data as a histogram to observe skewing. If possible, use a method which allows you to compare the distribution of price between vehicle class.
3. Visualise the dates of departure and arrival to observe gaps or trends in the data set.
4. Visualise the frequency of each fare type in the data to understand the spread within this category.

### Assess and Manage Missing values

1. Identify whether there are missing values in the Data
2. Which columns are they in?
3. Inspect some rows which contain them.
4. Drop all rows which have missing `vehicle\_class` and `price` and `fare`
5. Aggregate/ group or visualise the data by vehicle class and fare for further analysis. What does this tell us about how ticket price varies with respect to vehicle\_class and fare?
6. Fill the remaining missing price column values with the calculated mean of all the prices.
7. Check you have now removed any missing values from your data.

**De duplication**

1. Assess if the data contains any duplicated rows. As the dataset constitutes real ticket price search results, there’s a good chance duplication has come about due to the data collection method.
2. Consider what are valid duplicate records in this data and decide what to do with those – should we keep these? If yes, should a new unique identifier be added to the data?

### Assess Outliers

1. Identify outliers in the price column, preferably in an automated way.
2. Examine these outliers. Do they appear to be erroneous or is there a valid reason that they exist?

### Stretch exercise

1. As it appears price depends upon both vehicle\_class and fare, we may choose to replace missing price values with the average for their vehicle\_class and fare category, instead of the simpler mean imputation method used earlier. Write some code which does this- you will ideally define and populate a new column which can be analysed against the previous method applied, to pick the most suitable solution.
2. Perform some analysis which seeks to answer : is there a significant statistical relationship between destination (where the traveler is going) and fare (type, not price)? For example, is it reasonable to say that someone travelling from Madrid to Seville will typically travel on a promotional fare?