1. **Practical**

I am going to explain about which dataset I have chosen and what are the transformations I did to the datasetUsing Pentaho (which is a freely available software to use many of the ETL tools).

**Step 1:**

Downloading a dataset from Kaggle website make sure that you would organize those files in a specific folder , wher you can access easily , in my case I have downloaded a dataset containing the full information about different cars.

I didn’t find the dataset again after downloading it from the website, so I have uploaded in the drive and shared here below.

<https://drive.google.com/drive/folders/1ADqvdfrR5JTclPtiH8OryfTh82d0nElT?usp=drive_link>

**Step 2:**

In Pentaho, A dataset should be taken as input initially , it should be done with respect to its extension also. Accessing the right file with right extension plays a major role, so be careful in mentioning / browsing the path and give a name.

In my case both the datasets were in the form of text , so I took two different csv file input in the design section, because I had two datasets.

**Step 3:**

To sort a particular dataset by rows, we use sort rows , which we take the input thorugh pipeline and sort it in ascending order .

I mapped each dataset to a different sort rows through connecting them via piplines, weher pipelines shpw the flow data. It is very important to map the pipelines very well.

**Step 4:**

In order to merge two datasets and sort them together , we use sorted merge, generally it can sort any number of datasets into a single dataset.

I mapped both the sorted datasets to the sorted merge, in order to get every data sorted in ascending in one file.

**Step 5:**

If we want to remove some duplicate data , we use unique rows, which helps to display unique data and remove duplicate ones.

I had mapped the sorted merge to unique rows , and also I had selected the company column to get only unique car company details.

**Step 6:**

If we want to remove some rows containing specific values , we use filter rows .

I mapped the unique rows to filter rows to remove only Electric cars and CVT in the column Transmission Type, so that I can view all the cars except only electric and cvt, by adding the logical conditions like(AND,OR and NOT) and selecting the column name and entering the values we want to filter out is the major part.

**Step 7:**

To view particular sub data consisting of minimum columns we want , we use select values. So that we can able to get the only columns we want.

I mapped filter rows to select values, to get company ,model, horsepower, torque, transmission type, drivetrain, fuel economy, number of doors, price, model year range , engine type. So iam getting every data which I want except body type and number of cylinders which I didn’t select and write yes.

**Step 8:**

Finally , if we want that particular data to be stored as a text file , we use text file output.

I mapped select values to text file output, to get the data which is been required by me after the transformation will be stored in the folder where I specified it’s path.

So , This is the end of transformation what I did to the datasets .

My transformation is shared below:

<https://drive.google.com/drive/folders/1zHoHm9NilUbRVXyiuMMzBZbrBMY4-QtM?usp=drive_link>

Screenshot of My final completed transformation by running It:

