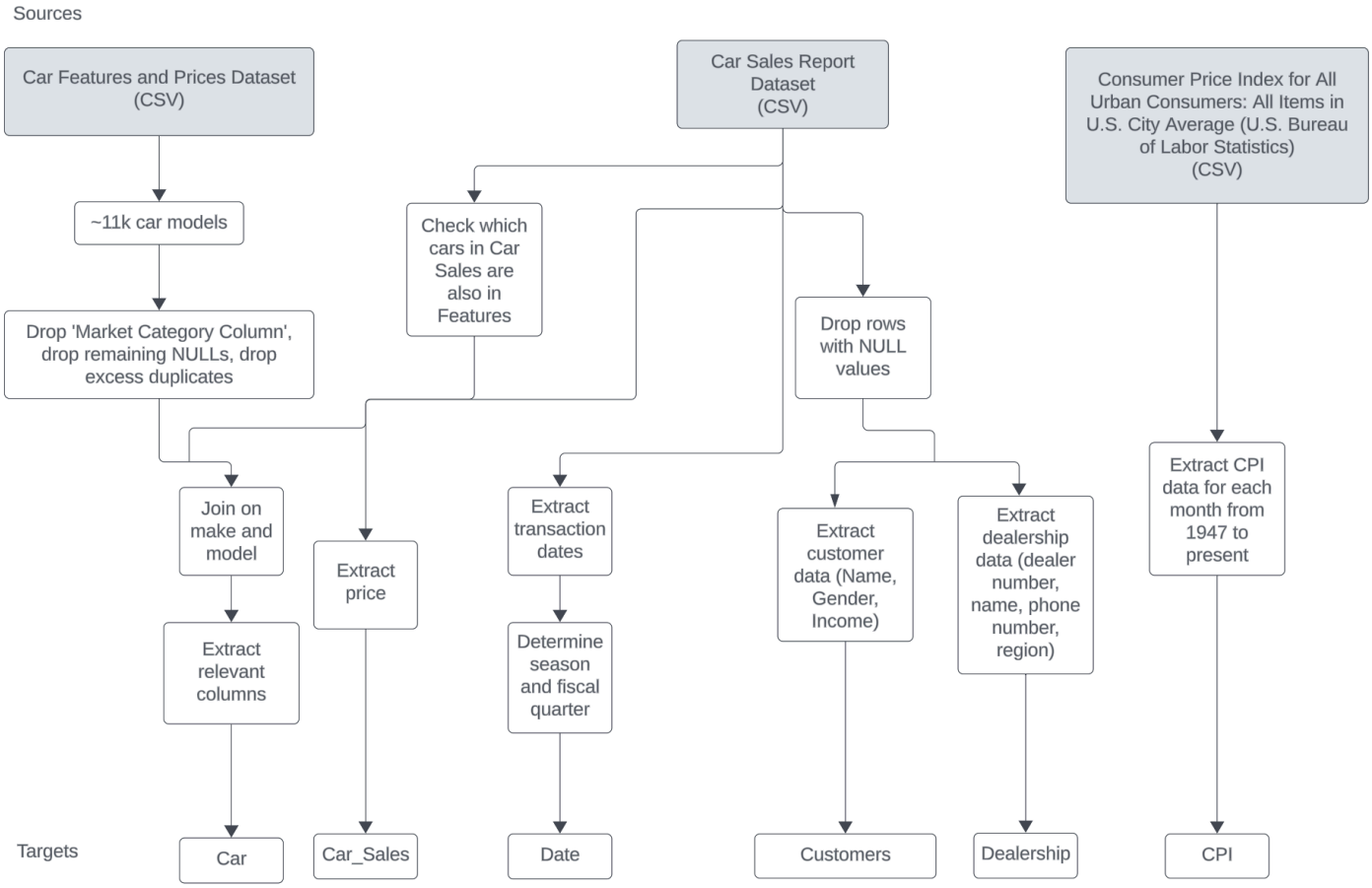
High Level Schematic:



Data Processing Challenges and Solutions:

Num Values

In the sales dataset, there was only one null value. Because this was a very insignificant number of affected rows, we decided to drop the row containing a null value.

In the features dataset, the “Market Category” column contained 3742 null values. This is a significant number of null values. We felt that meaningful analysis of our data could be performed without this column so we dropped the column. This dataset also had 69 and 30 null values in the engine horsepower and engine cylinder respectively. We decided to drop these rows because they would not result in considerable data loss.

Duplicate Values

The features dataset had 715 duplicate entries. Duplicate entries in this particular dataset meant the exact same car with the exact same features were entered multiple times. There is no need to keep the additional entries so we take the first instance of duplicates and drop the rest.

Joining Sales and Features Dataset

To populate our car table in the database, we used data from both the sales and features datasets and joined the two datasets on make and model. When joining the datasets, we noticed that there were several instances where one row of the sales dataset would match multiple rows of the features dataset. This meant that the joint dataset had many rows that were actually the same sale of a car but with every matching variation of features. We only want the features list to add information to each entry of the sales dataset so this duplication of rows was an unintended effect. Unfortunately, it was impossible for us to be able to tell which of the car trims in the features dataset was the one purchased by the customer in the sales dataset. We decided to aggregate the duplicate rows by taking the mean of numerical features and the mode of categorical features.
This join resulted in data loss of both the sales and features datasets as we performed an inner join so entries that could not be joined on make and model from either dataset were dropped.