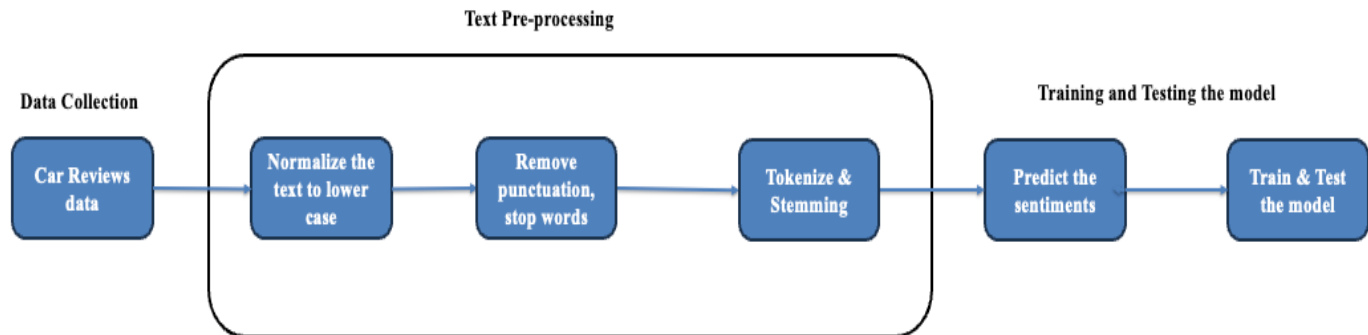# SENTIMENT ANALYSIS PROJECT:

Project Goal: To generate insights from customer feedback, enabling the rental agency to streamline their car inventory based on the sentiments extracted from each review.
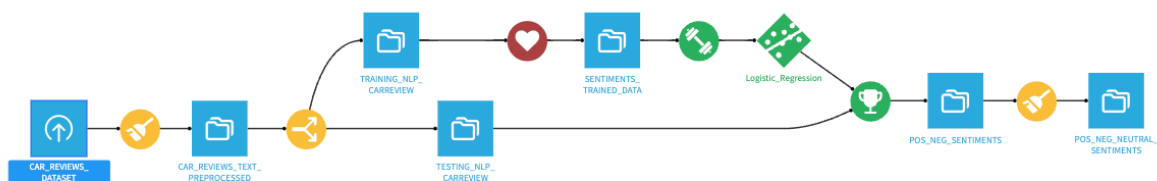
Project Flow:

There are 3 main steps:
1. Data Collection
2. Text Pre-processing
3. Training and Testing the model

Data Source: The dataset has been sourced from Kaggle, a popular platform for data science and machine learning projects.
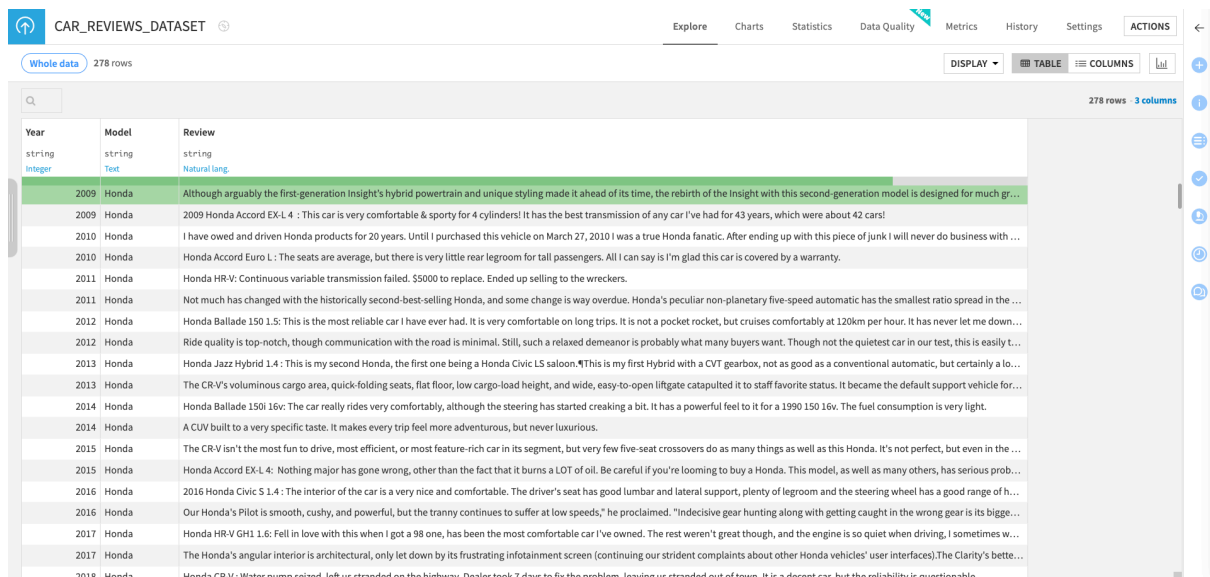
Features:

- **Year:** Year the car was made, providing temporal context to the sentiment analysis.
- **Model:** The model of the car being reviewed, allowing for analysis at a granular level.
- **Reviews:** This feature contains the textual customer reviews, which will be the primary data used to predict sentiments.

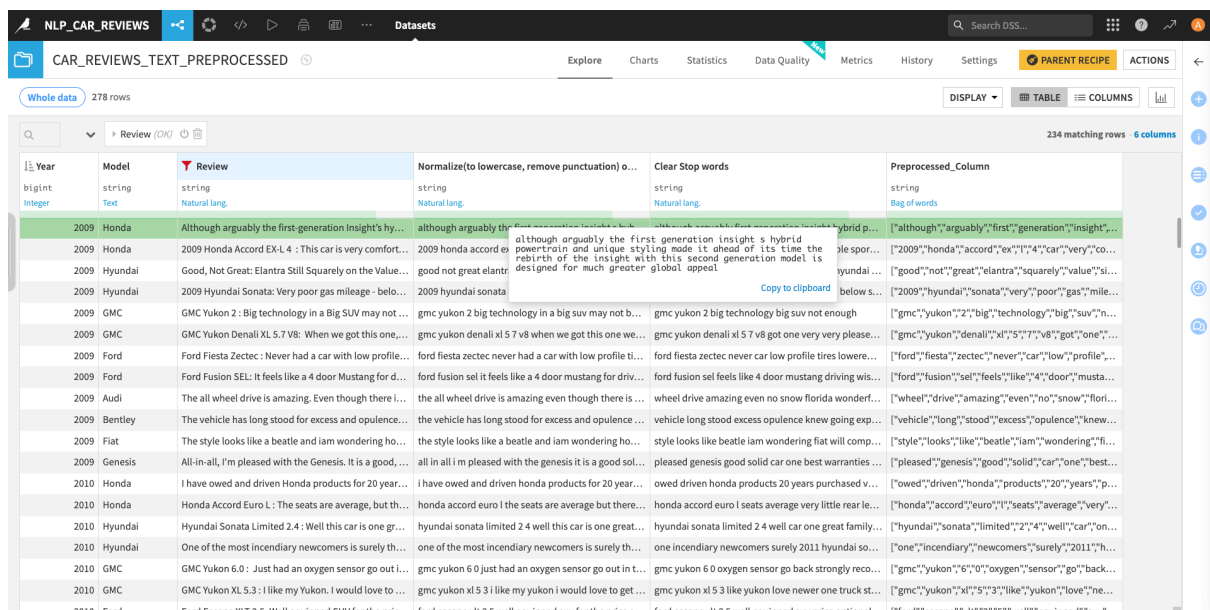

Sentiment Analysis Flow in Dataiku

Step1 : Load the Dataset in Dataiku

Step2 : Perform Text Pre-processing on the Review Column

1. Convert to Lower Case and remove punctuation

I/P: "The car was GREAT! It exceeded my expectations, and I had an awesome experience!!!"

- **Convert to Lower Case:**
  - Before: "The car was GREAT! It exceeded my expectations, and I had an awesome experience!!!"
  - After: "the car was great! it exceeded my expectations, and i had an awesome experience!!!"

- **Remove Punctuation:**
  - Before: "the car was great! it exceeded my expectations, and i had an awesome experience!!!"
  - After: "the car was great it exceeded my expectations and i had an awesome experience"

2. Clear stop words : To remove common words that do not carry significant meaning and are not useful for sentiment analysis. These words, known as stop words, include terms like "the," "is," "in," "and," etc. In python, we use NLP libraries such as NLTK

Before: "the car was great it exceeded my expectations and i had an awesome experience"
After: "car great exceeded expectations awesome experience

3. Stemming and Tokenizing :

Stemming : Reduce words to their base or root form, which helps in consolidating different variations of the same word.
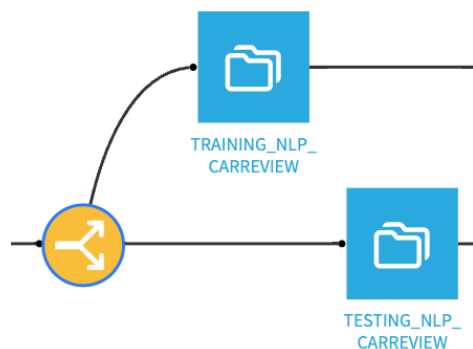For example, "running," "runs," and "ran" can all be reduced to "run."

Tokenizing : To split the text into individual words or tokens, which is a crucial step for further analysis.
Example: "The car was great" becomes ["the", "car", "was", "great"]

Pre-processed Tokenized Review column

Then, Splitted the data into two parts: 70% for training and 30% for testing.

- **Training Data:** The subset of data used to train machine learning models, enabling them to learn patterns and relationships from the input features and their corresponding labels.

- **Testing Data:** The subset of data held out from the training process, used to evaluate the trained model's performance on unseen examples to assess its predictive accuracy and generalization ability.

I used the Sentiment Analysis Plug-in in Dataiku to predict sentiments on training data. Since the training data doesn't have corresponding labels such as positive or negative, I explored its functionalities and learned how to effectively apply it within the platform.
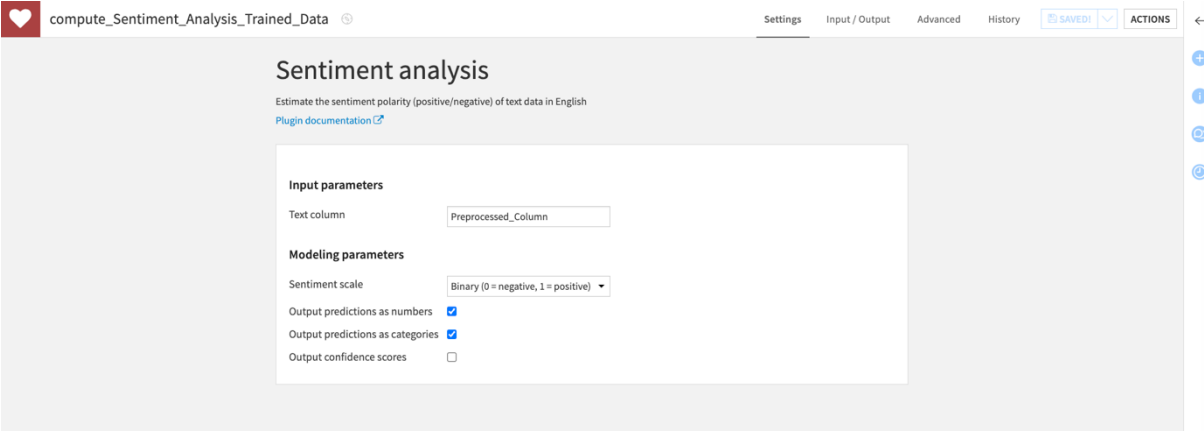
Then, I predicted sentiments on the training data. The image below highlights the sentiment predicted as Positive.

Deployed the Logistic regression model and trained on Sentiments_Trained_Data, then used this model to test on Unseen Testing Data
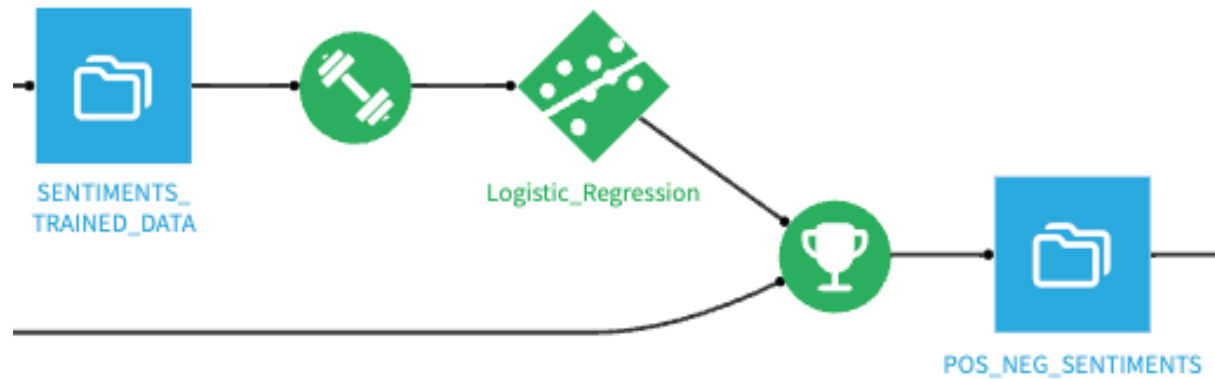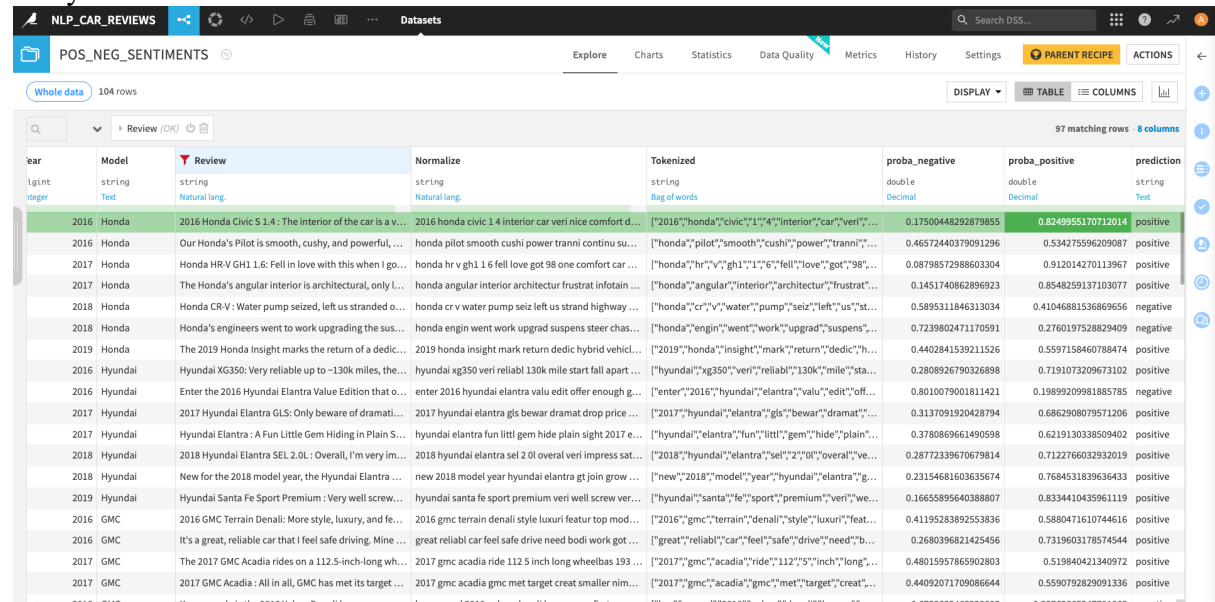

In the initial iteration, the model achieved an accuracy of 76%. I further assessed its performance using additional metrics, such as the confusion matrix, detailed below:
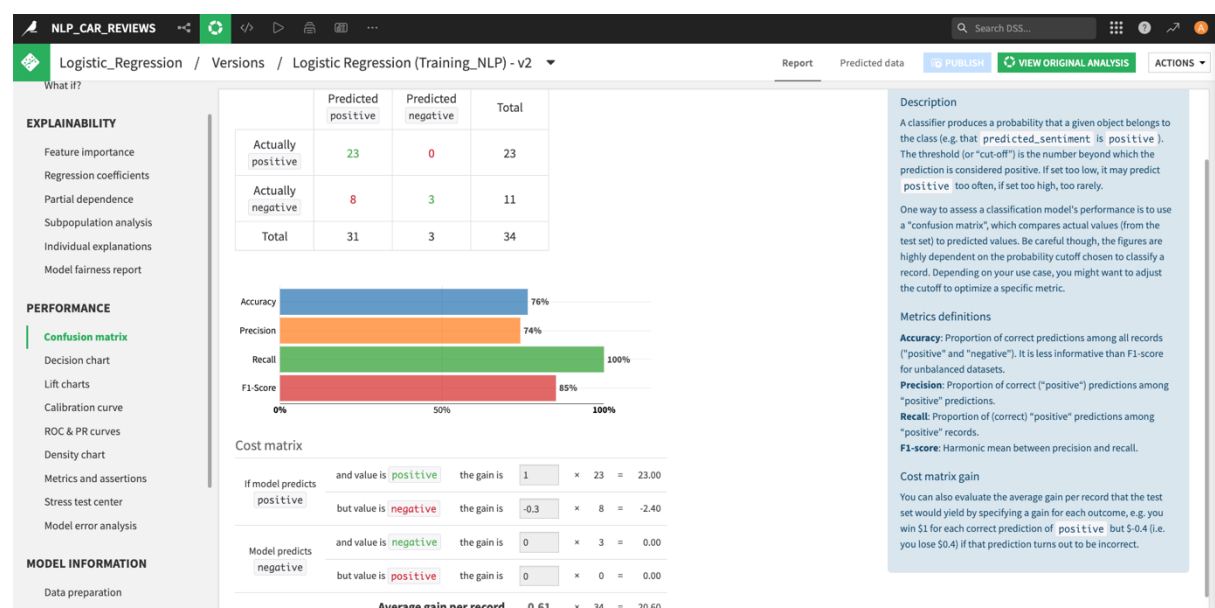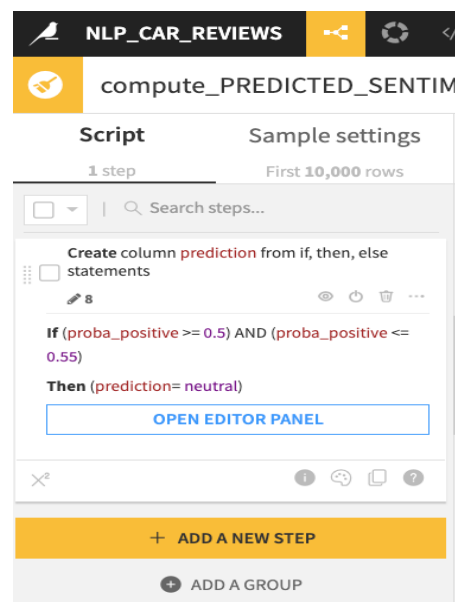
Positive and Negative Sentiments:

These below image represent whether reviews are seen as positive or negative based on the analysis.

Added another condition in prepare recipe that if the given polarity is between 0.5 and 0.55 then the sentiment is Neutral

Predicted Sentiments :

The below image displays the final predicted sentiments for each row, corresponding to the customer reviews analyzed.