Step by step explanation

- Filtering the data:
  - There were many empty cells in the given csv files.
  - So, I remove some rows with empty values in the columns where at most 7% data was missing.
  - But there were some columns with 15% or 25%+ missing cells.
  - For some of these columns I used ***ffill*** to fill up the cells. Because, sometimes if consecutive columns consist of same values, then in the datafile only 1st column remain filled.
  - Now from review_title column I observed the maximum data consists of a year and some statement inside bracket.
  - I extract these two, as I thought this can be important for the data.
  - I ignored some columns like user_name, review_description etc.
- Applying ML algorithms:
  - This was a multi-class classification problem.
  - Initially I transform all the categorical data by labelling.
  - Then I split the train data again in two parts test and train.
  - Then, I apply 5 ML algos:
    1. Decision Tree
    2. Random Forest
    3. Logistic Regression
    4. Naïve Bayes
    5. KNN (for k=5, 8, 2)
  - The best result I got in the **Random Forest.**
  - But the accuracy rate was not so good, `63.62543414373497%`.
- Observation:
  - The main observation I've made that **I should include the review_description in the analysis.**
  - It can be like this, some rapidly used word can be picked from the description and depending upon those words on or more columns can be added. I think then a better accuracy can be achieved.
- Final Note:
  - I've added *variety_new* column in the test data, which is the classified value using Random Forest algo. This is in the file named, Updated_test_data.csv.
- GitHub link: https://github.com/nondeterministicNilu/Niladri