Data Cleaning for Decantlo –

* Checked if there are any nan/missing values in data
* Checked histogram plots to see the range of values, unique values to see if there is anything incorrect or out of context
* Changed the column type of price to the appropriate type (from object to float64)
* For country, filled the missing values with ‘unknown’
* For abv values, dropped rows which had abv>100 (as abv takes values btw 0 and 100),

Checked the total number of missing values in abv and found that it was very less (<1%), so dropped these rows

* For missing values in year, we saw that the % was just 5%. So decided to keep these rows to use the other column values for further analysis of the data. This can be removed when fitting a model with year

Feature Engineering –

* We calculated the age of the wine from the year
* Created a new column ‘score’ which will have the rating value only if the number of reviews is>5 (putting threshold as 5 to avoid bias)
* As size for almost all bottles was 75cl, imputed the missing sizes with the most frequent value
* As sizes were different, calculated prices of all wines proportionally according to size 75cl (as this was the common size) – scaledprice
* To address the skewness in price data, we calculate the log price
* Since the unique country values are many, added a new column for country codes for cleaner visualisations

Things to do on merged data-

* # # one hot encoding of wine\_type (after merge)

# pd.get\_dummies(decantalo, drop\_first=True, columns=["wine\_type"])

* Can drop year with NA values (based on %)