Exploratory Data Analysis Report

Author:

João Vitor de Paiva Marcotti

2 years of experience in data science/analysis roles. Currently working as data analyst at FIESC and as data science freelancer at Whiskystats.

Database:

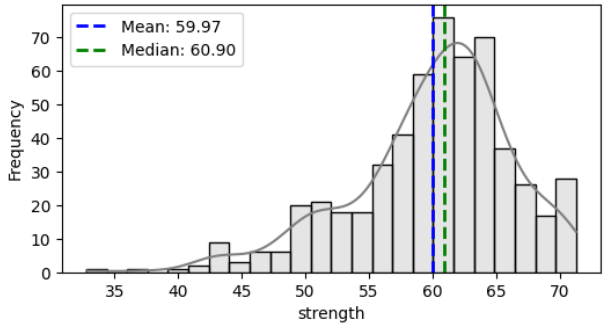
Whisky Casks Auction Database

Available at Kaggle: https://www.kaggle.com/datasets/joaopaivaa/whisky-casks-auction-database

1. X variables exploration

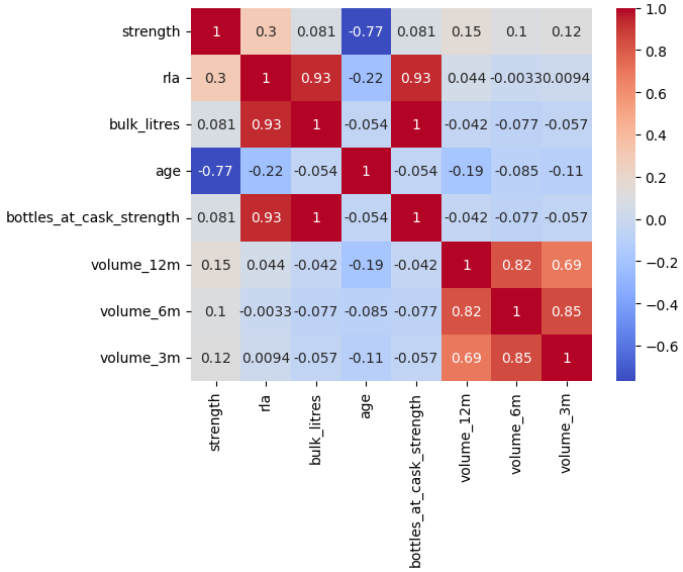
1.1 Numerical X variables

The numerical X variables were analyzed looking at its histograms, as well as its mean and median values.



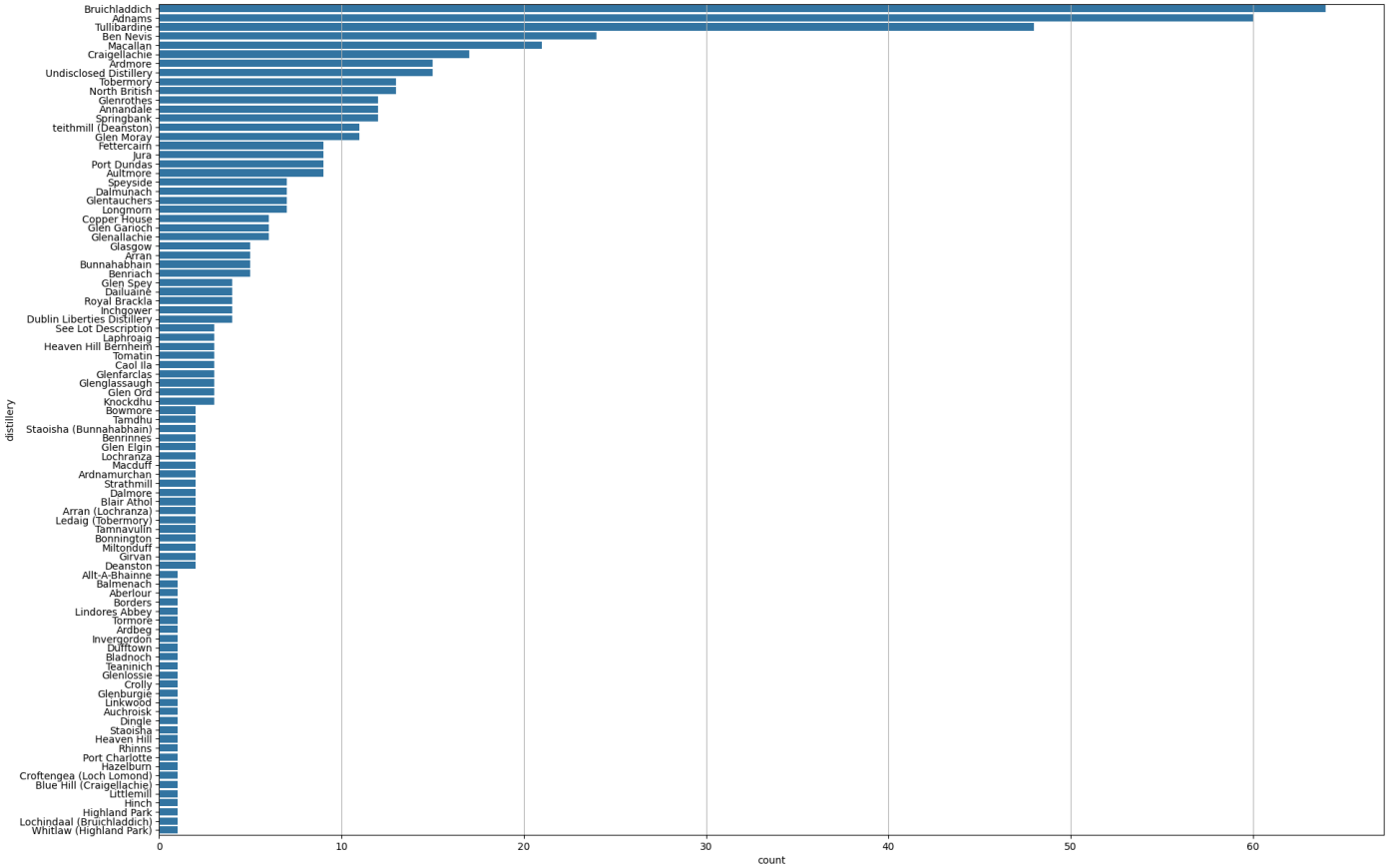
The correlation between these variables was also analyzed, and was found to be of high value between:

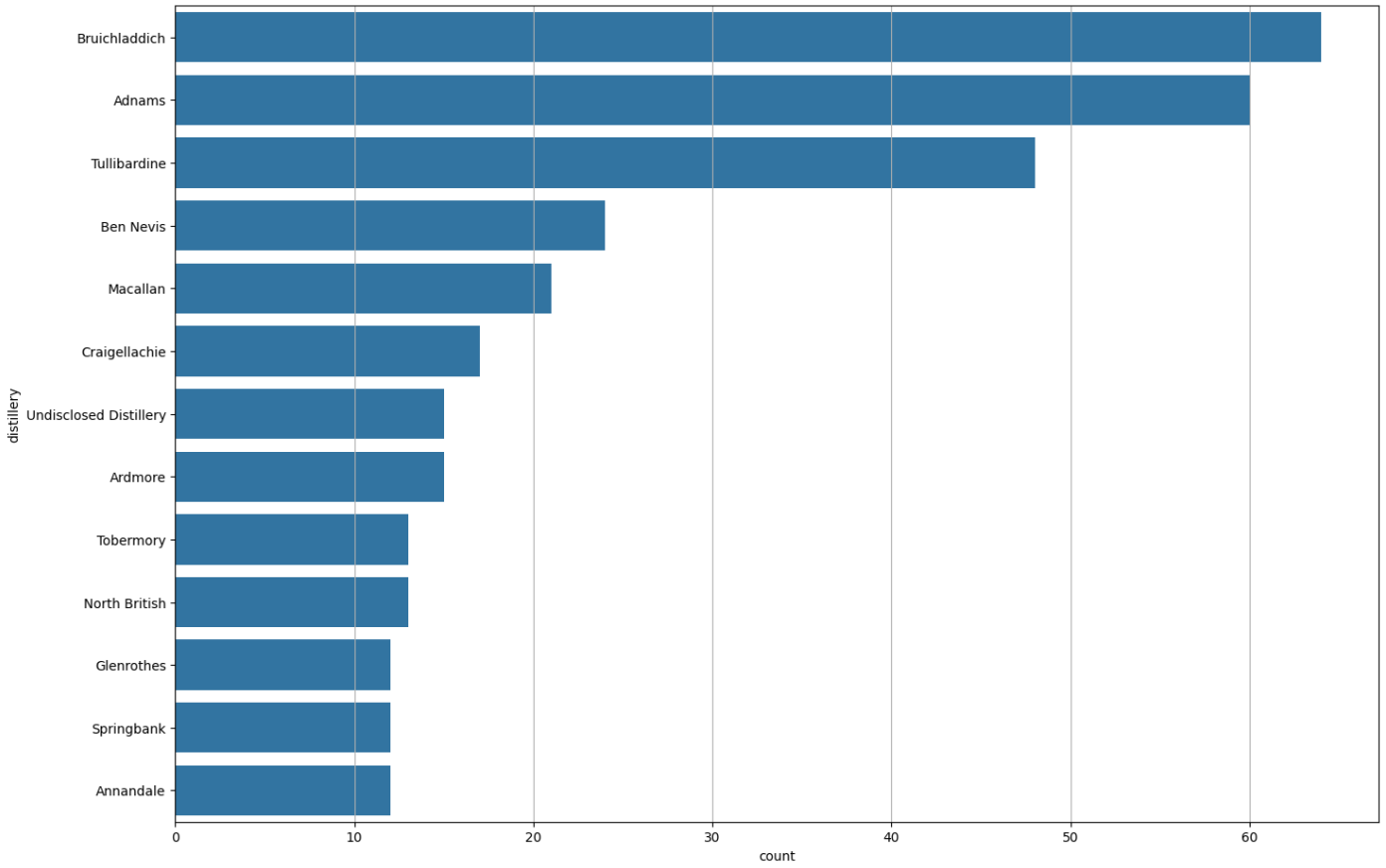
- Strength and age (inverse).
- RLA and bulk litres and bottles at cask strength.
- Volume 3 months, 6 months and 12 months.



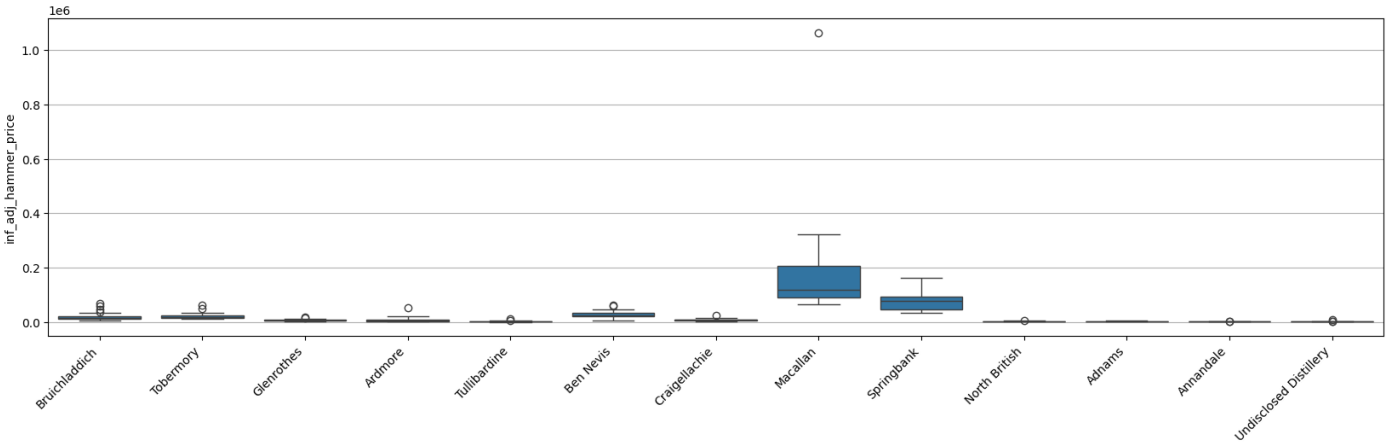
1.2 Categorical X variables

Some variables like distillery have a high number of categories. Therefore, to reduce its cardinality, the categories with at least 2% of the total values were selected, regarding the inflation adjusted hammer price Y variable.

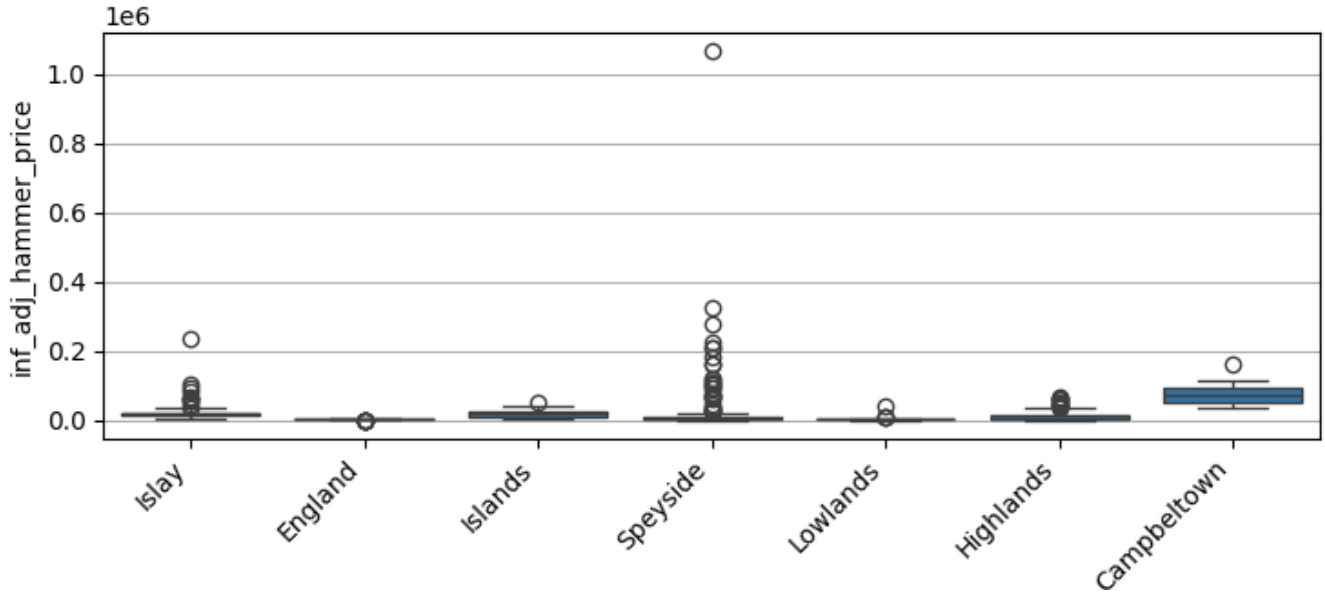




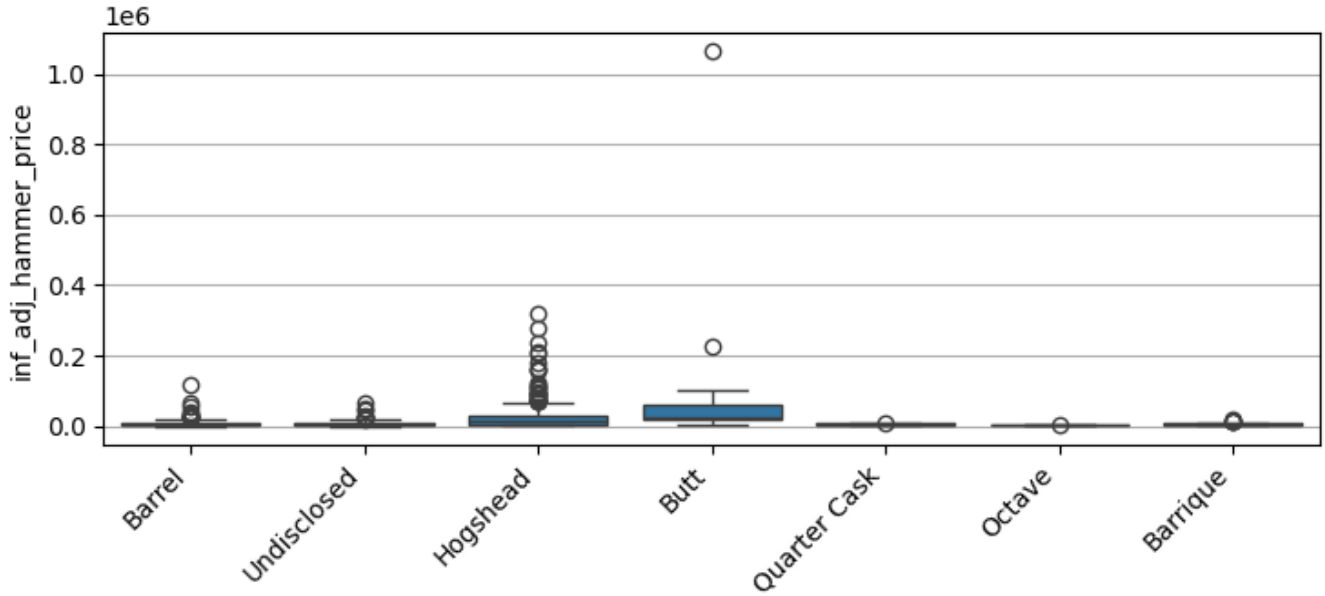
Their box plot distribution was also analyzed, searching for categories with an average Y variable higher/lower than the others.



Regarding the distilleries, Macallan and Springbank have an average price higher than the other distilleries.



Regarding the region, Campbeltown has an average price higher than the other regions.



Regarding the cask type, hogshead and butt have an average price slightly higher than the other types.

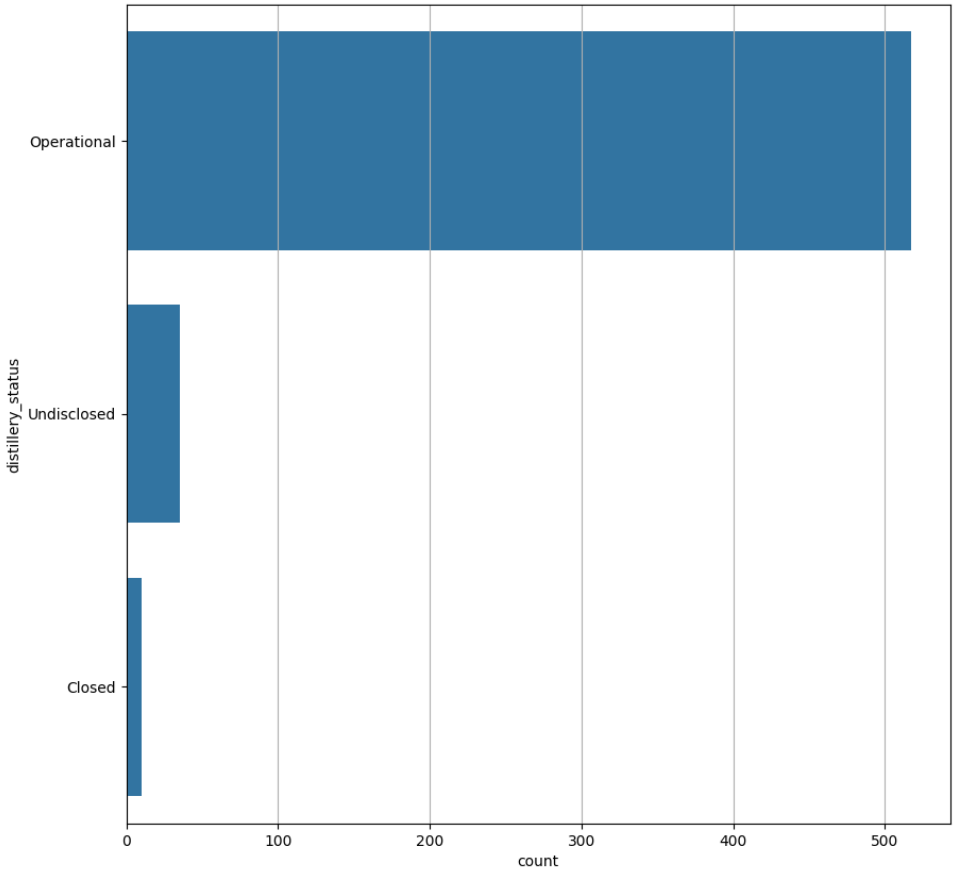
The same patterns happen to the other 3 Y variables.

1.3 Data cleaning

The NaN values were replaced by “Undisclosed” in the following columns:

- distillery.
- region.
- country.
- distillery_status.
- cask_type.
- cask_filling.
- previous_spirit.

Columns country and distillery status were removed because of having unbalanced categories. For example, the value “Operational” appears over 500 times in the distillery status column.



The Y variable inflation adjusted hammer price per bottle at cask strength was removed of the analysis because of having a similar meaning, price per volume, as inflation adjusted hammer price per litre of alcohol.

Casks with strength below 40% ABV were removed, as they no longer meet the legal criteria to be classified as whisky.

2. Y variables normalization

Multiple Y variables were tested:

- Inflation adjusted hammer price
- Inflation adjusted hammer price per litre of alcohol
- Inflation adjusted hammer price per age
- Inflation adjusted hammer price per litre of alcohol per age

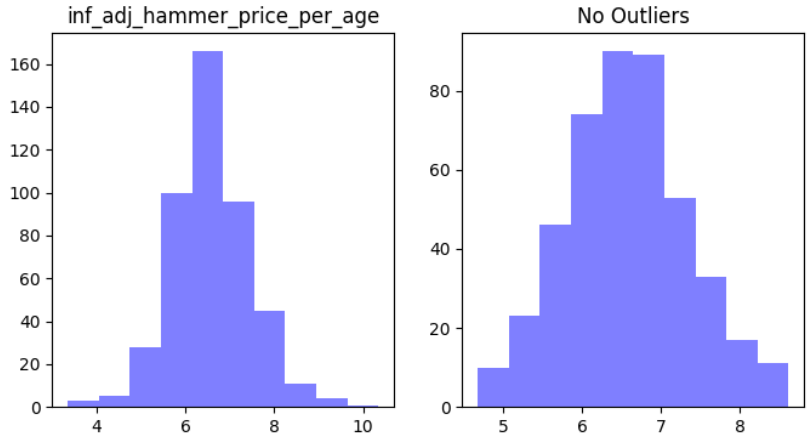
Log and Box-Cox transformations were applied to evaluate the best normalization for the Y variables, and evaluated through the Shapiro-Wilk test.

Shapiro-Wilk test logic:

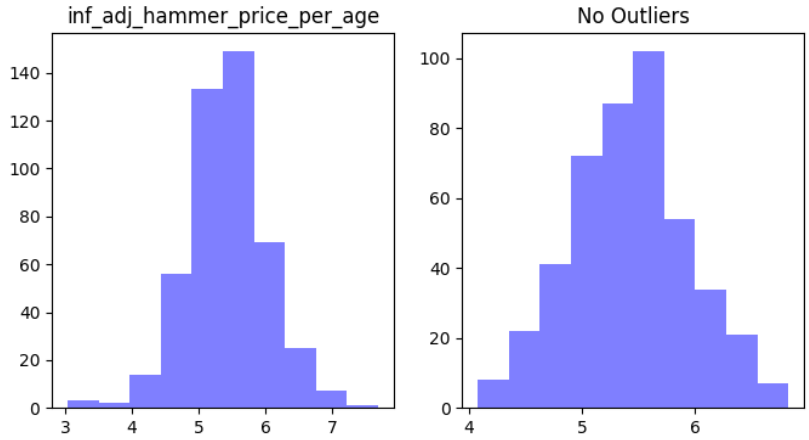
P-value	**Interpretation**	**Conclusion**
P > 0.05	Fail to reject H0	Data is **normally distributed**.
P <= 0.05	Reject H0	Data is **not normally distributed**.

Inflation adjusted hammer price per age showed the best results after normalization, achieving P-values of 0.14 and 0.48 for log and box-cox, respectively.

Log normalization for inflation adjusted hammer price per age:



Box-cox normalization for inflation adjusted hammer price per age:



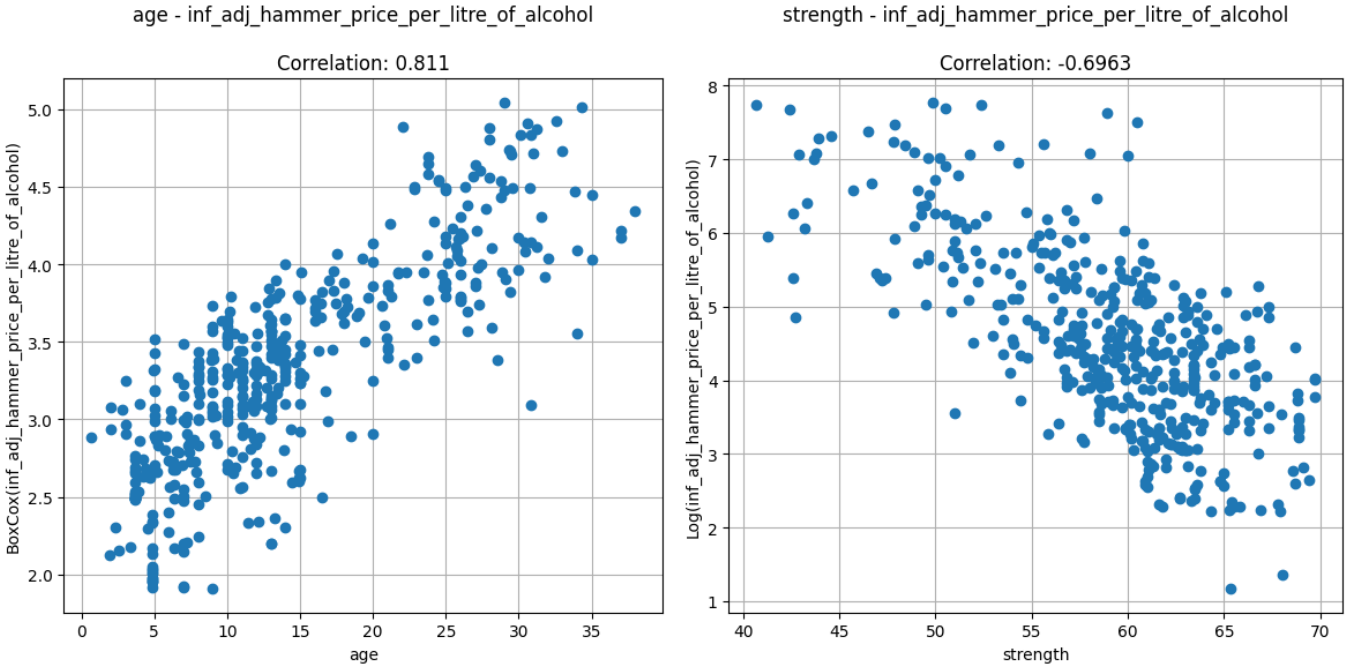
3. Relationship between Y and X variables

Correlations between all numeric X variables and the four Y variables were evaluated. The table displays the maximum correlation value for each X variable.

y column	x column	y transformation	x transformation	correlation
inf_adj_hammer_price_per_litre_of_alcohol	age	Box-Cox	None	0.8110
inf_adj_hammer_price_per_litre_of_alcohol	strength	Log	None	-0.6963
inf_adj_hammer_price_per_litre_of_alcohol_per_age	rla	Box-Cox	Box-Cox	-0.4063
inf_adj_hammer_price_per_age	bottles_at_cask_strength	None	None	0.3334
inf_adj_hammer_price_per_age	bulk_litres	None	None	0.3334
inf_adj_hammer_price_per_litre_of_alcohol	volume_12m	Log	Log	-0.3012
inf_adj_hammer_price	volume_6m	Log	Log	-0.1534
inf_adj_hammer_price	volume_3m	Log	None	-0.1005

Even not being considered a normal distribution according to the Shapiro-Walk test, inflation adjusted hammer price per litre of alcohol is the only analyzed Y variable with high correlation (above 0.5) in relation to at least one X variable.

Age and strength showed the highest correlation with this Y variable normalized through Box-Cox and Log transformations:



4. Features selection

Between the numeric X variables, just age and strength were selected, because they are the only variables of this type with high correlation with at least one of the Y variables analyzed.

Distillery, region, cask type, cask filling and previous spirit were selected between the categorical X variables, as their categories are not unbalanced.

Moreover, as inflation adjusted hammer price per litre of alcohol is the only analyzed Y variable with high correlation (above 0.5) in relation to at least one X variable, it's the only one selected.

Therefore, the final database, intended to be used to train a machine learning model, consists of the following columns:

- Age.
- Strength.
- Distillery.
- Region.
- Cask type.
- Cask filling.
- Previous spirit.
- Inflation adjusted hammer price per litre of alcohol.