Arcadian Contumacious, 22 Years: Likely an Expensive Whisky

Naina Wodon

Brown University DSI

https://github.com/nwodon/Whiskey-NLP

1. Introduction

Whiskey is a popular beverage consumed in the United States, with a number of different styles and types. Due to the large diversification of the whiskey market, the beverage can be sold at many price points.

However, it is unclear what exactly determines the price of a whiskey. The age and rating of a whiskey may impact the price, but what is the effect of brand name? Does a brand have an impact on price? Similarly, do whiskey flavors such as vanilla, or whiskey processes such as the type of wood cask the whiskey is aged, in have an impact?

I hope to determine what features are predictive of whiskey price using a dataset called *2,2k_ Scotch Whiskey Reviews* which was obtained from Kaggle, originally from Whiskey Advocate, and natural language processing (NLP) to gain new insights about the whiskey market.

The original dataset of 2,247 observations contained the following variables: ID, name, category, review score, price, currency, and description. All whisky prices were in $US dollars and so currency and ID were dropped. Of the remaining features, name and description were text features, category was a categorical feature with 5 whisky types, and review score and price were continuous features. The overall dataset had 8,655 features, most of which were generated using NLP.

2. EDA

Extensive exploratory data analysis was conducted and many plots were generated. The key insights are presented below.

First, I examined the distribution of whisky price. In figure 1 below, price is highly right skewed with one whisky costing $US 157,000. The distribution was normalized by taking the log-price (figure 2).

Figure 1

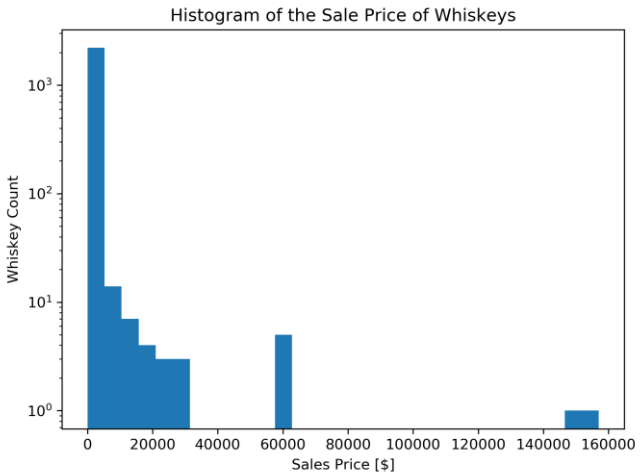


Figure 2

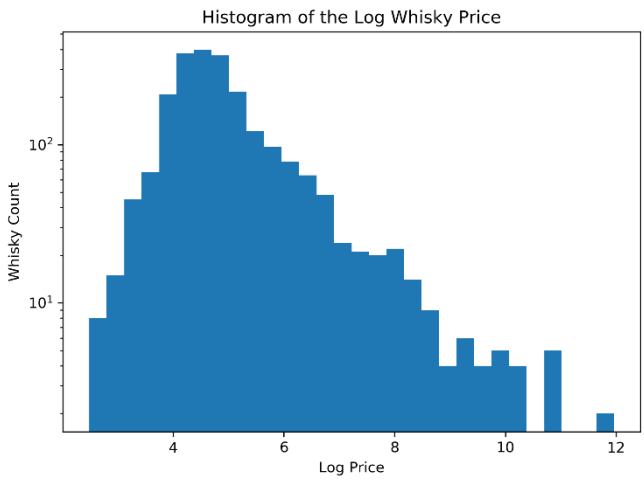
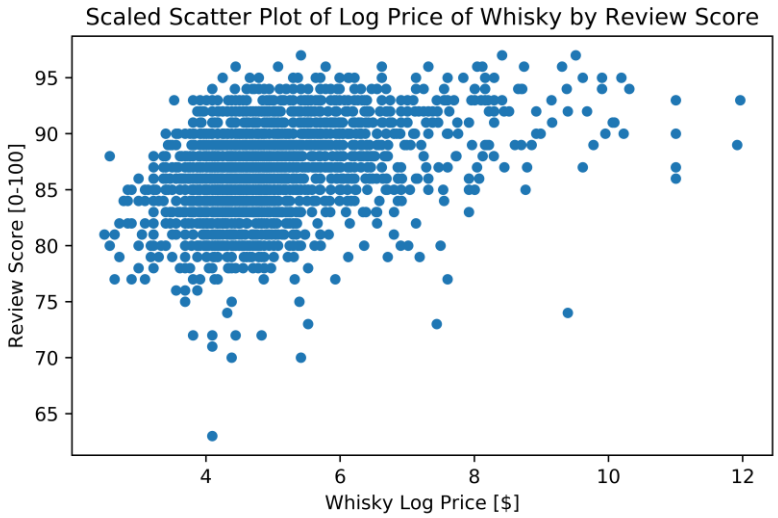


Figure 3 presents a scatterplot of log-price and review score. The features are slightly positively correlated, but given that the relationship was small, a new feature, review score squared was generated.

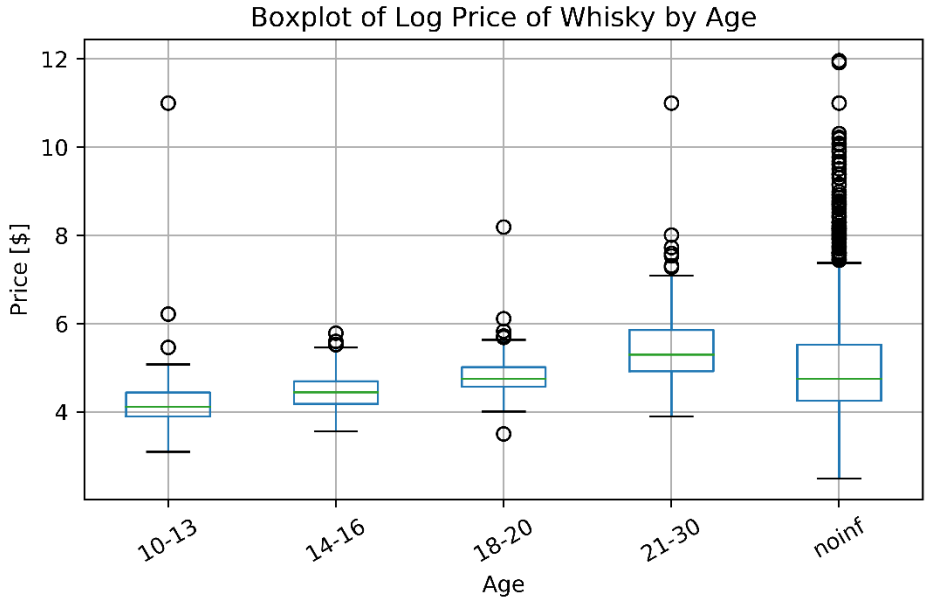
Figure 3



The bulk of this report involves extracted text features, the processing of which will be covered in the methods section. The age, brand, and most frequently occurring words were extracted using NLP. The extracted features were visualized in different ways.

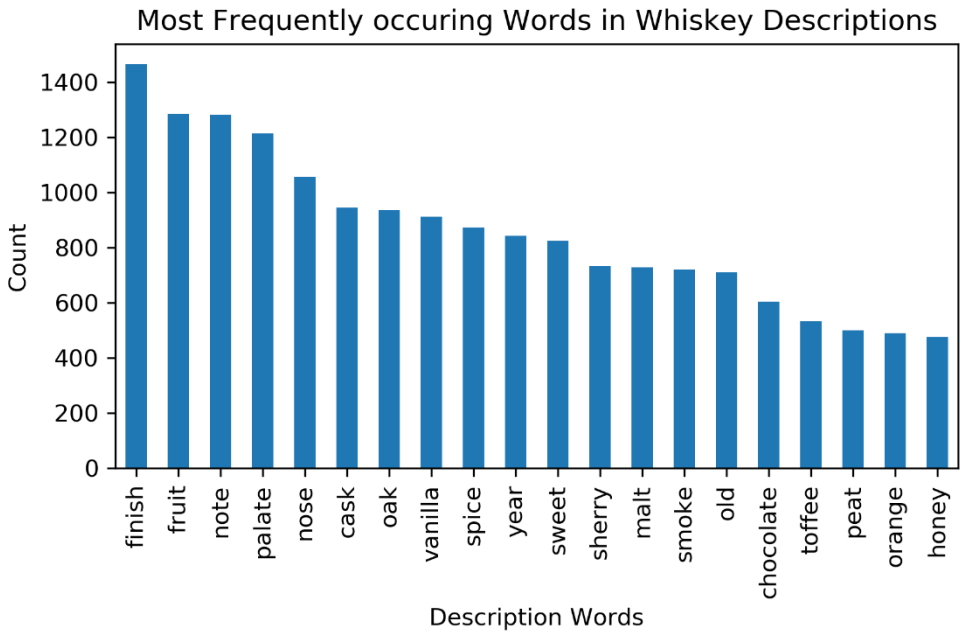
Whiskey specialists agree that the older a whisky, generally, the rarer and the higher quality the whisky (GQ). Thus, I visualized the relationship between age and whisky log-price (figure 4). There is an increasing relationship between age and average log-price. There is a large jump between the categories 18-20 years, and 21-30 years (relatively common among high-end whiskies). The category 'noinf' represents the whiskies with no information and accounts for 60% the observations.

Figure 4



The most frequently occurring words in whisky descriptions are plotted below (figure 5). The top words are whisky-specific language like finish and note which are used to describe the common whisky flavors vanilla, and spice. Other common words like cask and oak relate to manufacturing processes.

Figure 5



3. Methods

A total of 77 features were included in the final model. From the original data, the category, review score and review score squared were included. Through Natural language processing whisky age, brand, alcohol percentage, the number of words in the whisky name, the average number of characters per word in the name, the number of words in the whisky description, and the average number of characters per word in the description (7 features) were extracted and included. An additional 8,645 features corresponding to the frequency counts per observation of the word (excluding stopwords) were generated from the description. Of these features, the 67 most frequently occurring words were incorporated in the final ML models.

The whisky brand and age features were generated by removing the stopwords and then iterating through every name, creating a document term frequency (DTM) where the rows represent the whisky observations, and the columns represent all the terms separated by white spaces. To extract the age, the numbers 10-30 were extracted from the DTM and searched through the original excel file to ensure the numbers actually corresponded to whiskey age. 17 and 26 did not correspond to age and so were removed. Ages were split into categories based on the available literature on whisky (GQ). To determine which whisky brands to extract, two different articles listing the best 50 whiskies were consulted (D'Marge and Men's Journal). The whiskies in the list were converted into strings where one word represents the full name. For example, the whisky Johnnie Walker was condensed to Walker. The matching features names from the DTM were extracted and represent the frequency of a whisky brand. Next, the features were cleaned to ensure that for an observation there were not two ages or brands (i.e. Johnnie Walker the Royal Route corresponds to the brand Walker, not Royal) and then combined into one brand feature with 19 categories and one age feature with 5 categories.

The alcohol percentage feature was generated using regex. The word count and average character count per word features were generated in a similar manner. To generate the frequency of the description words, the text was first cleaned and then lemmatized.

The i.i.d. data was preprocessed within my lasso, random forest (RF) and k-nearest neighbors' regression (KNN) pipelines to avoid data leakage. The categorical features age, brand and category were one hot encoded and all the other features which were continuous were scaled using standard scaler. This preprocessing was fit and transformed on the training set (X_other) which was further split into 5 training sets where one set was trained at a time and the test set was transformed.

The models were built step by step, adding additional features at each stage, but we will only discuss the final models. To run the lasso and RF regressions, a K-fold Cross Validation (CV) pipeline was developed to train the model and tune the hyperparameter(s) at the same time while avoiding overfitting the data.

Lasso regression was selected as the supervised ML model because it can be used for variable selection and this was ideal given the large number of features. The alpha hyperparameter was tuned on a log-space between 10^(-5) and 10^(5) (intervals of 10). RF regression was also implemented to determine if it performed better than a linear model. The hyperparameters tuned were the min. sample split with values between 5 and 150 (intervals of 10) and the max depth between 1 and 30 (intervals of 5). The range for max depth was adjusted to be between 20 and 40 (to avoid edge values). For the KNN model, the number of neighbors was tuned with values between 5 and 500 (intervals of 50).

The R^2 score was used to evaluate the performance of the different models because it can be interpreted easily. If the R^2 score is 1, all the variation in log-price is explained by the model (a perfect predictor). To measure the uncertainties due to the splitting each model was tuned using 10 different random states where for each state a new hyperparameter was chosen. For the 10 states, the mean R^2 test score and standard deviation of the different R^2 scores were calculated and outputted in the table below.

Table 1. Summary of Model Performance

Supervised ML Model	Mean R^2 Score	Standard Deviation
Lasso Regression	0.3760	0.0397
Random Forest	0.3929	0.0489
K Nearest Neighbors	0.1756	0.0354

The alpha hyperparameter for the lasso regression picked a value of 0.0001 most frequently, although sometimes 0.001 appeared. For the RF hyperparameters, the max depth split was most frequently 20 and the minimum sample split was most commonly 10. All the neighbor's values were 50.

4. Results

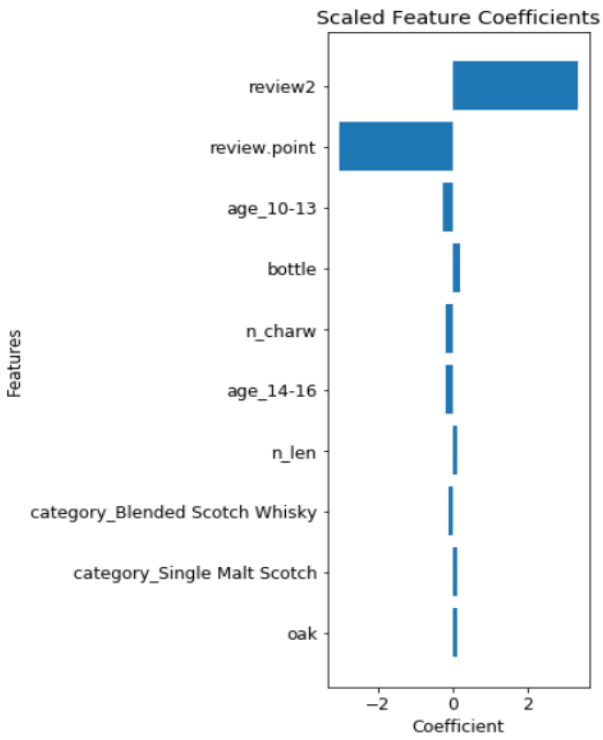
The KNN regression performs quite poorly with a mean R^2 score of .17 because KNN does not perform well with high dimensional data. The lasso and RF models performed significantly better and were within 1 standard deviation of each other making it difficult to determine the better model. However, given that lasso regression is more interpretable, it was chosen as the final model.

The mean R^2 score for the lasso regression was .376 and can be interpreted as the model explains 37.6% of the variation in the log-price. The model performed much better than the baseline R^2 of 0 which represents the mean log-price. Permutation feature importance (figure 6) and scaled coefficients of the lasso regression (figure 7) were

used to calculate global feature importance. Both models highlighted the same features as important, but examining coefficients of the model, shows that age groups 10-13 and 14-16 years is negatively correlated with the log-price. This dummy is what drives the importance of the age feature in the permutation model. From the permutation plot, it is clear that the review score and review score squared drive the log-price model as shuffling those observations results in a negative R^2 scores. Only the word bottle appears in both plots and is positively correlated with log-price. Words like finish and oak appear in only one of the plots. Surprisingly, in both plots the number of characters per word and the number of words in the whisky name are quite significant.

Figure 6

Figure 7



This information is useful to whisky distillers as they can draw several conclusions about the feature importance. As expected, younger whiskies tend to be cheaper as age is related quality. Whiskies aged in cask oaks tend to be slightly more expensive and including the word bottle in a whisky description is correlated with a higher price. Whiskies with more complex names tend to command higher prices. This may be because expensive whisky will state that it has been aged for many years and have a complex yet memorable name. The most important predictor of price is review score and this signals to distillers that if their whisky is highly rated, they can use this to market and sell their bottle at a premium. It is also meaningful to note that the brand of whisky has a startling impact on whisky. When these values are shuffled, the R^2 score drops down to 0 signaling that the brand impacts the log price. While this may seem obvious, it's interesting because brand prestige, a function of marketing, is correlated with higher prices!

5. Outlook

To improve the model, the coefficients of the Lasso model could have been tested for statistical significance such that statistically significant features remained. The adjusted R^2 score penalizes the model for adding insignificant features. Many features in the final model had insignificant feature importance values so removing them would improve the R^2 score and result in a simpler model.

Furthermore, whisky brands included in the brands feature and the age-groups cutoffs in the age feature were based on outside literature. Whisky articles were consulted because they provided useful insights and cutoffs to transform the data. However, the data may not match the cutoffs used and so choosing different cutoffs may result in a different model. Moreover, the NLP techniques were only able to extract age and brand information for 40% and 30% of the data respectively. With more time, the observations that were coded as no information could be manually checked and this would likely boost the explanatory power of the model given that the age of the whisky had a quite high feature importance.

When shuffled, many of the word features had no impact on the R^2 score. If the features selected had been based on term-frequency inverse-document frequency, an additional method which determines the relative importance of a word, the model would improve. Support vector regressions and XGBoost were not included in the analysis because of their long training times. These can be implemented in the future.

In terms of additional data, adding another feature which categorized where the whisky was distilled may increase the model performance as some distilleries are famous for producing expensive whiskies. Similarly, adding data about customer perception of whisky quality can improve the model as perception is correlated with price (D'Marge).

6. References

50 Best Whiskey Brands In The World. D'MARGE, 19 Sept. 2019, https://www.dmarge.com/2019/02/whiskey-brands.html.

Ando, Koki. "2,2k Scotch Whisky Reviews." *Kaggle*, Whisky Advocate, 13 June 2018, https://www.kaggle.com/koki25ando/22000-scotch-whisky-reviews.

Compton, Natalie B. *Here's Why Your Whiskey's Age Matters*. GQ, 11 Dec. 2017, https://www.gq.com/story/why-your-whiskeys-age-matters.

Editors, The. *Best Whiskey: The 50 Best Whiskeys in the World*. Men's Journal, 6 Nov. 2019, https://www.mensjournal.com/food-drink/the-50-best-whiskeys-in-the-world-w211382/monkey-shoulder-blended-malt-scotch-w211406/.

Other Kaggle projects relating to the dataset pertained to sentiment analysis and clustering:

- https://www.kaggle.com/koki25ando/cluster-analysis-of-whisky-reviews-using-k-means
- https://www.kaggle.com/gsdeepakkumar/classy-whisky-approach-through-nlp
- https://www.kaggle.com/arindamgot/whisky-reviews-sentiment-analysis-n-gram-lda
- https://www.kaggle.com/samarthagarwal23/scotch-recommendation-using-universal-sentence-enc
- https://www.kaggle.com/theenduser/scotch-whiskey-data-cleaning
- https://www.kaggle.com/bsivavenu/review-of-best-whisky