BOSTON UNIVERSITY **Boston University** Questrom School of Business

BA820: Unsupervised Machine Learning
Prof. Brock Tibert – 11.23.2022

Final Project Deliverable
Scotch NLP Analysis
Team A3

Team members
Rochel Chan, Rafikiel Seyvunde, Aditya Sinha, Weijia Suo, Alexander von Schwerdtner

1. Introduction

1.1 Project Mission

How can scotch reviews be used for new categorization to maximize profit for liquor store owners?

1.2 Business Problem (that motivates analysis)

During lockdown, the whiskey market was on fire as wealthy connoisseurs stayed home drinking and entertaining friends. As the spirit category continues to grow, so will the demand for blended Whiskey at premium price points. According to market analysis, factors responsible for this growth include the need for "premiumization and innovation". Therefore, our dataset of scotch reviews will identify, using NLP, methods to determine the quality of scotch.

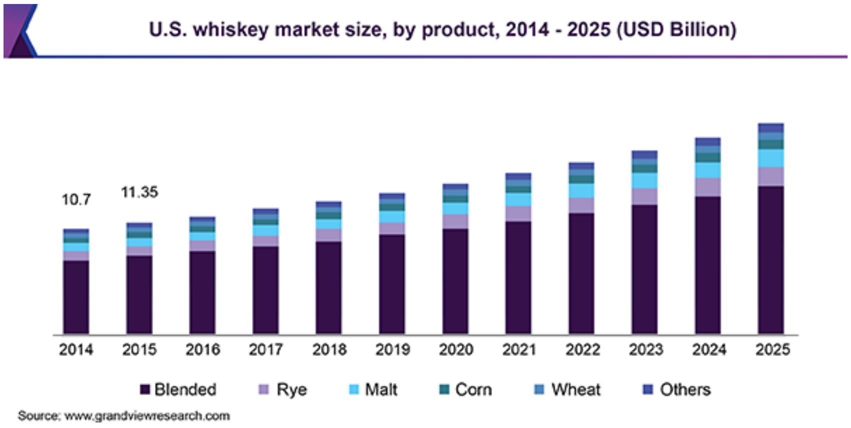
Ultimately, the goal of this project is to maximize profits for the liquor store owner and differentiate the owner from the competition within the industry in terms of innovative reorganization and marketing strategy.

We will also strive to build a category prediction ML model for new scotch to assist the owner with maintaining the organization.

1.3 Whiskey market size

The global Whiskey market is expected to grow at $33 Billion 2027, with blended scotch being the fastest growing category in the market.

- CAGR of 6.8%
- Blended whiskey driving category demand


U.S. whiskey market size, by product, 2014 - 2025 (USD Billion)

Source: www.grandviewresearch.com

1.4 DataSet

<u>Data Source**:**</u> https://www.kaggle.com/neilcosgrove/scotch-whiskey-reviews-update-2020

BA820: Unsupervised Machine Learning
Prof. Brock Tibert – 11.23.2022

Original Source: Web scraping from *The Whiskey Advocate*

This is a compilation of Scotch Whisky Reviews from the "Whisky Advocate" taken as of January 2021 (so results are up to 2020) with 2,247 reviews.

Our dataset is retrieved from kaggle and user ‘thatdataanalyst’ has full credit for this dataset. This dataset contains reviews of Scotch whiskey and contains 2247 rows and 7 columns. To analyze the reviews, we will focus on the description column to perform text analysis.

Column Name	Description
id	ID number
name	Name of whiskey
category	Type of whiskey
review.point	Number of review points given to the whiskey
price	Floating point price
currency	Currency in which price is reported
description	Review comments

Note that the following scale is used to review whiskey for the ‘review.point’ column.

Numerical scale (points)	Qualitative review
95-100	Classic: a great whiskey
90-94	Outstanding: a whiskey of superior character and style
85-89	Very good: a whiskey with special qualities
80-84	Good: a solid, well-made whiskey
75-79	Mediocre: a drinkable whiskey that may have minor flaws
50-74	Not recommended

2. Exploratory Data Analysis

2.1 Data Cleaning & Preprocessing

After loading the dataset using pandas, we first conducted data exploration. Our dataset consists of 2247 rows and 7 columns. Each row represents an individual scotch and the columns represent various attributes of whiskey with customer review.

We started by looking at the column types. Id and review.point are int values and all the other columns are object type.

- For our analysis, we converted the price column to float type. We used regex, map, and lambda functions to change any string representation of price to float type.
- Renamed description.1.2247. to review_description.

Next we checked for any duplicates or null values.

- We found 39 null or NA values in review_descriptions. We decided to drop these values because we want to use our review_descriptions column for NLP, NLU, and clustering analysis.
- We found 57 duplicates for the name column. We decided to merge the duplicate name by taking average for numeric columns such as price and review.point and using join for string review_description column.

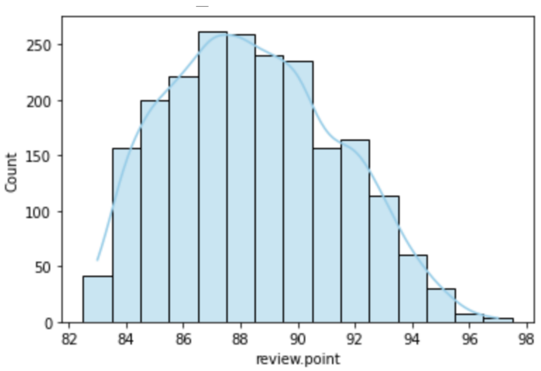
For feature engineering we extracted two columns.

- alcohol% - This float column was extracted from the name column using regex defining the percentage of alcohol for the given scotch bottle. We found there were 10 scotch that didn’t have alcohol percentage so we filled those with the mean for that column.
- price_per_point - This column was calculated by dividing price by review.point which will be helpful during our analysis.

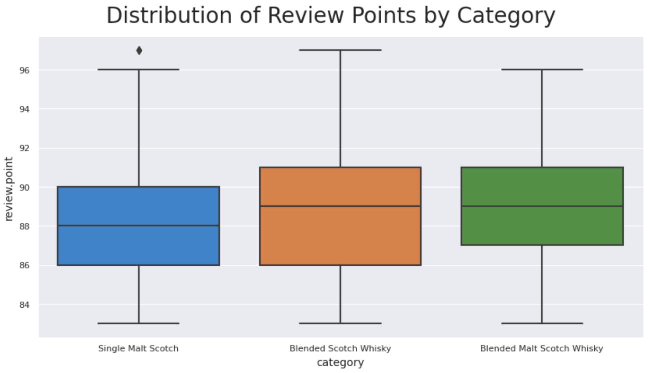
At last, we dropped redundant and unnecessary columns such as ‘Id’ and ‘currency’. When we completed the data cleaning process we managed to keep about 95% of the data. There are now 2151 scotch rows and 6 scotch feature columns with no missing and duplicate values.

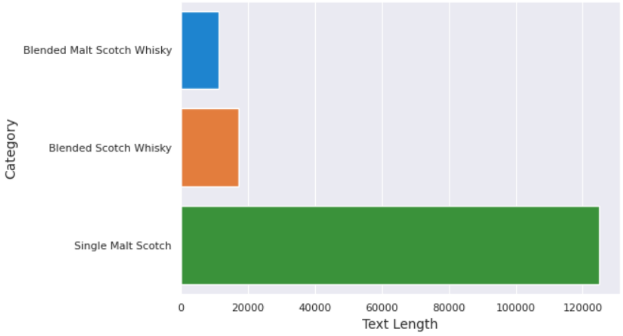
2.2 Visual Analysis

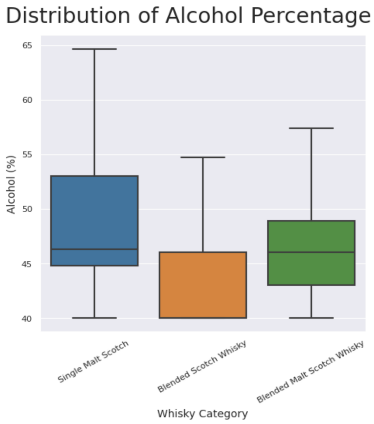
To better understand the distribution of review points and alcohol percentage, a histogram and boxplot was created respectively. The histogram shows that most whiskeys have a review point around the range 85 - 91. This means that most whiskeys are perceived as ‘very good’ and less as ‘outstanding’. Notice that none of the whiskeys received a review below 82 and very few of the whiskeys are reviewed as ‘classic’.



The distribution of alcohol percentage was categorized according to the three types of whiskeys (Single Malt Scotch, Blended Scotch Whiskey, and Blended Malt Scotch Whiskey) with outliers removed from the boxplot. . On average, the alcohol percentage for blended scotch whiskey is lower than the other two categories. Single Malt Scotch has the highest alcohol intensity and the range of all three varieties is about 5-8%.



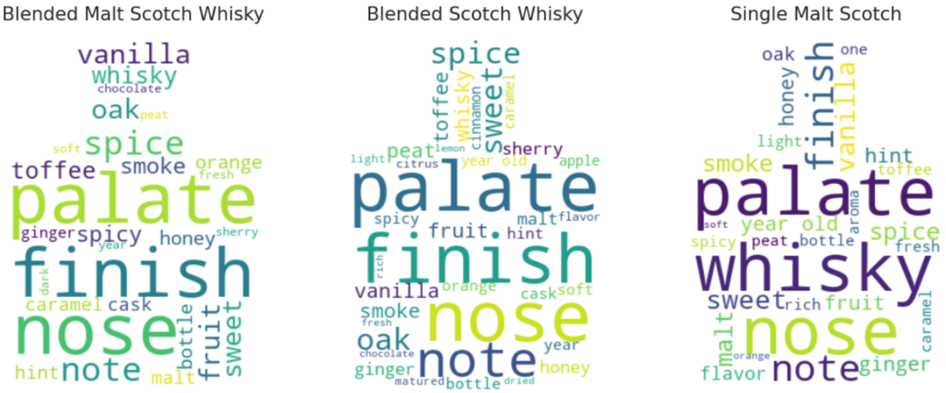




We analysed the description column using word cloud and bar chart. We grouped the dataset based on the category and combined the review_description for the three categories. Then calculated the length of combined review_description and used the word cloud to represent the most occurring word for each category.

Single Malt scotch has the highest review length, followed by Blended Scotch Whiskey and Blended Malt Scotch Whiskey.

Word cloud to look at the most occurring word for each three categories of scotch.



2.3 EDA Conclusion

After conducting the EDA for the Dataset we have ensured that we are able to use the dataset for our Machine Learning and that there are no issues with parts of the data.

3. Analytical Findings

3.1 Classification Prediction Model

We chose to run a prediction classification model for this business problem so that liquor store owners can re-run the model when they get new products or when they would like to regroup/re-market and predict which grouping it could belong to. They can use this methodology to adjust their business strategy or even just their marketing strategy by bundling together certain products.

3.2 NLP for Cluster Analysis

NLP to form clusters:

- Cleaning text for NLP
  - We used SnowballStemmer to combine/stem different forms of the same words to prevent getting various forms of words that basically have the same meaning.
  - Then we use RegexpTokenizer to tokenize and keep only words including those with apostrophes.
- Preprocess the text
  - TfidfVectorizer (option)
    - To get into the doc term representation
    - `tfidf = TfidfVectorizer(stop_words = 'english', tokenizer = RegexpTokenizer(r'[a-zA-Z\']+'))`
    - Finally, we got `[2151 rows x 5667 columns]` for vectorizer.
  - spaCy (better performance)
    - To complete the NLP task, we generally need to preprocess the text. Compared with TfidfVectorizer, Spacy can process more efficiently, thus improving the accuracy of classification prediction.
    - Making lower case, removing `stopwords` and `punctuations` using spacy built in function.
    - Finally, we got `[2151 rows x 2178 columns]` for vectorizer.
    - As we can see, compared with TfidfVectorizer, Spacy enables us to extract more important keywords, thus making cluster results and classification prediction results more accurate.

- Compress the dimensions with UMAP

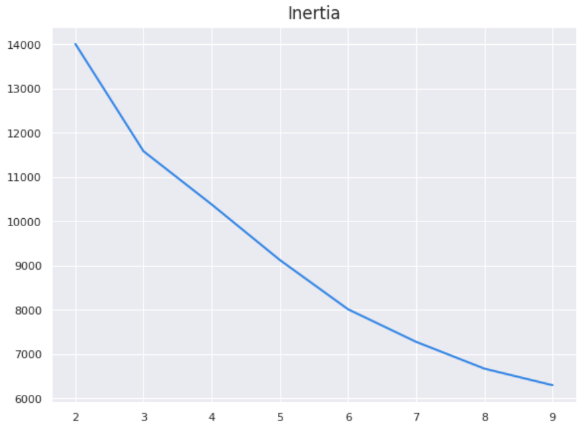
Clustering dataset

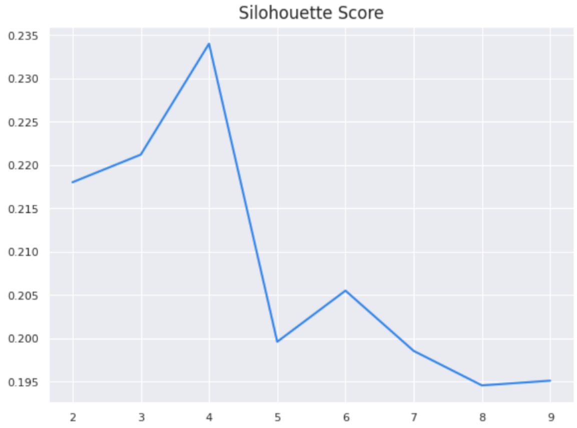
- Added e1, e2 dimensions from the UMAP to the cleaned dataset.

- Now using all these features to develop clusters.
- Convert categorical variable to numerical variable using one hot encoding.
- Before we conduct cluster analysis is to standardize the data. We rescaled featuresso that they all have a mean of 0 and a standard deviation of 1. This way, no feature has more influence on the clustering algorithm than the others.

3.3 K-Means

To choose an optimal number of groups, K-Mean clustering technique was performed. The graph on the left shows the decrease in inertia as the number of clusters increases. On the graph on the right, the blue line represents the silhouette score as clusters increase. At 5 clusters, we have the highest silhouette score and lowest inertia. To acquire the most accurate results, this would be the ideal number of clusters. We then calculated the number of datasets in each cluster and the mean of review point, price, alcohol, and price per point. Afterwards, we plotted this on a heatmap to explore the correlations. This will be further analyzed in our recommendations section.





3.4 XGBoost with hyperparameter tuning

We chose to use XGBoost from scikit-learn as it provides a large range of hyperparameters. We can leverage the maximum power of XGBoost by tuning its hyperparameters.

In order to evaluate the performance of our model, we needed to train it on a sample of the data and test it on another. We extracted the features and the target, our clusters from the K-Means, from our dataset and used the function train_test_split from scikit-learn to split the data into a test and train set (⅓ test, ⅔ train).

Before fitting the model we provided steps for the estimator by passing in the encoding with OrdinalEncoder and running the model with xgb.XGBClassifier and putting these into a Pipeline. We did hyperparameter training by passing the parameter grid into GridSearchCV with a 5-fold cross-validation. Finally we fit the model using the train data and calculated the accuracy score on the test dataset where we received an accuracy of about 98.59%.

```
steps = [("encoding", OrdinalEncoder()),
         ('model', xgb.XGBClassifier())]

pipe = Pipeline(steps)

param_grid = {
    'model__max_depth': [2, 3, 5, 7, 10],
    'model__n_estimators': [10, 100, 500],
}

grid = GridSearchCV(pipe, param_grid, cv=5, n_jobs=-1, scoring='accuracy')

grid.fit(X_train, y_train)
```

4. Conclusion & Recommendation

Our regression is run on all variables after we create clusters. In other words, clusters are based on the reviews and the predictions are based on all the variables.

Group 0 has the highest number of alcohols with 970 in total. They are the cheapest on average by price and the lowest by the review point.

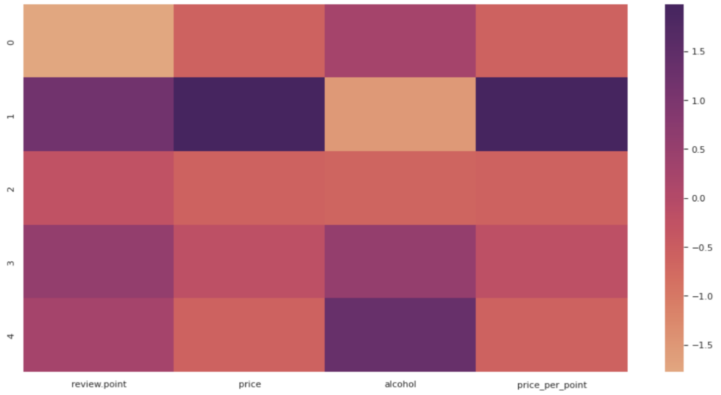
Group 1 has only one bottle and is the outlier with the highest review point and price. Therefore, it is in a separate cluster by itself.

Group 2 has 321 different alcohols and has the second lowest price per point and is in the middle range of scotch alcohol percentage.

Cluster 3 has 29 different alcohols and has high-class scotch within this cluster. It has the second highest alcohol percentage of 47.8% and has second highest review points as well.

Cluster 4 has 830 different alcohols and has the highest alcohol percentage with 49.86%. It also has a relatively high price per point.

The heatmap below shows the correlation between the five clusters and review point, price, alcohol percentage, and price per point. The darker the color, the higher the correlation.



To propose solutions to our initial business problem, we recommend the liquor store owner to have 2 main groups with the main differentiator as price. The first group will include cluster 1 and 3, with the highest priced whiskeys. The second group will include clusters 0, 2, and 4. The purpose of having two main groups is to optimize advertising resources and firm strategies. Having 2 main groups will be easier for the liquor store owner to target customers who are looking for higher priced whiskeys and customize marketing strategies accordingly. The goal is to maximize revenue by having whiskey groups while minimizing costs for the firm to better target their customers.

The implications of our project can be applied to liquor store owners who want to differentiate themselves from the competition. The liquor store owner could redesign its website based on the two groups by having separate pages for higher priced alcohols and lower priced alcohols. The company could then develop further marketing strategies to advertise these alcohols depending on the price point. For example, certain advertisements would better appeal to customers who are looking for premium whiskeys. The company could also perform A/B advertisement testings depending on the clusters to maximize conversion rates and increase sales.

For further implementation, the owner can also rerun algorithms as new data increases. The new model would help place new whiskeys within different clusters and potentially improve the prediction accuracy. In the long-term, this can help the liquor stores to better understand their customers' preferences and how to better advertise their products.