Whiskey Business

Team members: Patrick, Chris, Dylan

Overview

The goal for our project is to provide reviews and pricing information to consumers in state controlled liquor areas. As a proof-of-concept we are first starting with a single state and one alcohol type, whiskey. We will match state inventory with reviews and pricing sourced online into a user friendly dashboard.

Current Data Sources

- VA state liquor inventory: import CSV from online
- Reddit Whiskey Review Archive: import CSV from online
- Meta-Critic Whiskey Review: import CSV from online
- Proof66 Pricing and Reviews: scrape site for now, API integration in future builds

Current Data Process

Our functioning process has 5 steps:

1. Pull data down from online & scrape Proof66
2. Use Python to scrub data and perform matching
    a. Reason: Python gives us the best tools to scrub our data as well as the Fuzzy Wuzzy package for fuzzy string matching
3. Upload to HDFS and load into Hive
    a. Reason: HDFS will allow us to scale our tools to other states and liquor types without needing to migrate to a different architecture
4. Use Hive to build complete table of bottles and reviews
5. Import table to Google Data Studio for visualization and UI
    a. Reason: Data Studio is free and has multiple integration option. For now we will use a CSV but could transition to an API if we were to have an always-on server

Work-to-date and Initial Problems

Our largest issue is that there is no good primary key to match bottles across databases. We have leverage Python to scrub the names in our data allowing us to increase our match rate from 10% to 50% in the past few weeks. We are now starting to use Fuzzy Wuzzy for fuzzy string matching to increase our rate. Unfortunately many bottles are uniquely identified by a single word resulting in high fuzzywuzzy scores for poor matches. Currently we are testing different scoring parameters and minimum scores for a successful match. Increasing match rate will be one of our two major working areas.

We have currently loaded our VA and Reddit dataset into Hive and begun to build our final table. The second area of major work will be increasing the review data included in our final output. Reddit data represents the rating of the average consumer. Meta-critic allows us to include professional critic scores. Proof66 gives us the average rating based on competitions as well as national pricing. If we can successfully integrate all 4 databases we will give consumers the ability to rank bottles based on whiskey type, consumer rating (Reddit), critic rating (Meta-critic), or competition rating (Proof66), as well as price vs national benchmark (Proof66).

Conclusion

We have identified our data sources and started to work on creating a common key for pairing. Match rates have been our largest issue but we have found success with scrubbing rules and Fuzzy Wuzzy. Our current data process allows us to scale our process to all states and all liquor types without needing to change architecture. We are about 1 week out from a complete end-to-end process. After we have confirmed our process we will spend the remainder of the time focusing on our two work areas: 1. Increasing match rate 2. Additional database integration.