Capstone 1 - Project Proposal

Jeff Johannsen

BACKGROUND and MOTIVATION

Colorado is a hub for craft beer enthusiasts and this is one of the reasons I originally moved here. Knowing more about the quality of the beer and breweries in the local area and around the world will allow me to make better recommendations on what breweries people should visit and what beers they should try. This fits into a larger goal of creating a better tool for people to find fun activities in their local areas.

QUESTIONS

1. Colorado is considered “The Napa Valley of Beer”. Is this an accurate statement? Does the review data back up the claim that Colorado produces some of the highest quality beer in the country?
2. India Pale Ales (IPA) are the most popular type of beer produced by craft breweries today. Is this focus in line with the review data? How does the review data compare between different beer types?

DATA

My base data is a csv of craft beer reviews from the website BeerAdvocate. The data was provided through Kaggle: Beer Reviews CSV 172MB. I hope to expand upon this dataset by adding data from other ratings locations via web scraping and api usage.

This data contains 1,586,614 rows and 13 columns of data. There are 4 object cols, 3 int cols and 6 float cols. 10 out of the 13 columns are 100% non-null and the other 3 are very close to 100% non-null. Additional data added will increase the non-null rates considerably though. The datatypes of columns seem reasonable based on what they contain. Multiple composite columns will need to be added though.

Minimum Viable Product

- EDA using Pandas and Matplotlib to answer the questions above.
- Use functions and objects to make the analysis pipeline applicable to other datasets.

MVP+

- Add data from other data sources and review sites.
- Explore how these sources compare to the base dataset.

MVP++

- Expand questioning and analysis into tourism, restaurant, and bar industry data.