Project Report

INNOPOLIS UNIVERSITY

Beer Reviews

Students	Abdulmueez Emiola
	Ozioma Okonicha
Course	Big Data Technologies and Analytics
Semester	Spring
Year	2023

Contents

Note: Follow this Moodle link for more info on CRISP-DM shared by Armen

1. Introduction

Craft beer in general has been growing and becoming more popular over the last years; hence beer lovers from everywhere are always searching for more and more kinds of beers to try out. In order to fulfill their desires and quench their thirst, breweries are testing out various ingredients and new methods of brewing. This has resulted in several options, flavors and styles of beers. Now that there are so many options at the finger tips of the consumer, it can be difficult for them to decide on which beer they want to try.

In our report, we will use the Beer Reviews dataset gotten from BeerAdvocate, we will analyze the dataset in order to gain insights from this magnificent world of craft beer. The dataset contains over 1.5million reviews for different beers and is a rich knowledge base that we will use to understand what the consumers prefer, which beers styles are popular and even highlight future patterns in the beer industry.

Our analysis will be focused more on the exploration of how different beer attributes, for example style, abv and overall rating relate to one another. It will also touch on how these factors contribute to consumer satisfaction. When we identify the beer attributes that consumers see as the most important, brewers can then adjust their products in a way that fits the demand of their target audience. This was their sales and customer satisfaction will ultimately increase.

According to a recent report by the Brewers Association, in 2020, the craft beer market in just the United States is valued at a price over $29.3 billion. This value represents a 6% increase in comparison with the previous year, and this continues to grow even more each year [5]. Additionally, another report by Technavio says that the global beer market is expected to increase by a value of $97.1 billion between 2020 and 2024 [6]. The report also mentions two factors that are contributing to the growth of the global beer market as: demand for craft beer and the rise of microbreweries. Although, the report also notes that the beer industry faces challenges such as changing consumer preferences, government regulations, and competition from other alcoholic and non-alcoholic beverages.

Figure 1: Global Beer Market 2020-2024

From our choice of project, we hope to show the value that using data trends to make decisions can have in the beer industry. We also try to provide insights that brewers can make use of to change or choose their marketing strategies and product development. These insights could even potentially help future research that will be done in the field of craft beer. As the global beer market continues to grow and evolve, it is important that breweries stay ahead of their competition and one way to do this by understanding their consumers' preferences. Our analysis of the Beer Reviews dataset in this project will definitely offer opportunities since one can gain insights into the industry that will eventually lead to business success.

2. Business Understanding

Breweries are continuously seeking for methods to set themselves apart from the competition and draw in new customers in the fiercely competitive craft beer market. Breweries must comprehend their target market in order to develop products that appeal to them and succeed in this sector. By utilizing the Beer Reviews dataset from BeerAdvocate, we want to solve the business problem of determining the elements that influence consumer happiness in the craft beer sector. We want to learn crucial information about customer preferences for various beer characteristics, such as style, ABV, and overall rating, by studying this dataset. Breweries can utilize this information to produce new products, enhance current ones, and establish more efficient marketing plans.

To evaluate the current scenario, we will examine the current market for consumer behavior in the craft beer business and identify major trends and industry concerns. In order to find patterns and relationships between various beer qualities and customer happiness, we will then examine the Beer Reviews dataset using data mining techniques. Finally, we will create a preliminary plan that explains the procedures required to reach our goals and provide ideas other breweries wishing to better their goods and marketing tactics.

2.1 Current situation assessment

We researched existing literature and industry studies to better understand the trends and issues affecting the craft beer sector before assessing the situation there as it stands today.

According to the Brewers Association (2021) [5], the craft beer market in the United States continues to grow, with over 8,764 craft breweries operating in the country in 2020. However, the industry faces challenges such as changing consumer preferences, government regulations, and competition from other alcoholic and non-alcoholic beverages.

According to Technavio's research on the global beer market (2020), the market is expanding due to the development of microbreweries and the increased demand for craft beer [6]. The study found that because regional craft beers have unique flavors and styles, consumers are becoming more interested in them. The study also highlights challenges such changes in the price of raw materials and the effects of government regulations on the industry.

Also, we analyzed the Beer Reviews dataset from BeerAdvocate to get some insights on consumer preferences for various beer attributes. The dataset has over 1.5 million beer reviews and this serves as an information base that can be used to understand consumer behavior in the craft beer industry.

Overall, our current situation assessment sheds light on the importance of understanding consumer preferences and trends in the craft beer industry. Because this helps breweries to stay ahead of their competition. From analyzing the Beer Reviews dataset, our goal is to discover important patterns in consumer preferences for different

beer attributes, such as beer style, ABV, and overall rating. These patterns can then be used by breweries to develop new products and marketing strategies.

2.1.1 Inventory of resources*[Core]*

To complete the project, we have the following resources available:

- Personnel:
  - Business experts: While our team does not include members with experience in the craft beer industry who can provide valuable insights into industry trends and consumer behavior, we have our Professor Armen Beklaryan who has general business experience and whom we can approach with questions.
  - Data experts: Our team includes members with knowledge in data management and data analysis.
  - Technical support: We have access to technical support in person of our supportive Teaching Assistant, Firas Jolha, to help with any issues that may arise during the project.
  - Data mining experts: Our team includes members with some... experience in data mining and machine learning who can help with the analysis of the Beer Reviews dataset.
- Data:
  - Beer Reviews dataset: The dataset contains over 1.5 million beer reviews. The reviews have information about the beer's name, brewery, style, ABV, and user reviews.
- Computing resources:
  - Hardware platforms: We have access to high-performance computing resources (8GB RAM with HDP successfully installed) to support the analysis of the Beer Reviews dataset.
- Software:
  - Data mining tools: We will use various data mining tools, such as Python, and libraries such as Pandas, PostgreSQL, Hive, Sqoop, and PySpark, to analyze the Beer Reviews dataset.
  - Other relevant software: We will also use Streamlit for data visualization and reporting, including dashboards and charts.

With these above-mentioned tools, we have everything we need to finish the project and reach our goals of figuring out what parts of the craft beer industry may affect consumer satisfaction.

2.1.2 Requirements, assumptions and constraints

Requirements:

- Schedule of completion: The project must be completed within approximately 8 weeks, with progress checkpoints almost weekly on Thursdays labs.
- Comprehensibility and quality of results: The results of the analysis must be clearly presented and easily understandable by non-technical stakeholders. The quality of the analysis must be high, with rigorous statistical methods used to ensure the validity of the results.

Assumptions:

- The Beer Reviews dataset is representative of consumer behavior in the craft beer industry.
- The Beer Reviews dataset is accurate and complete.
- Consumer preferences for different beer attributes remain relatively stable over time.

Constraints:

- Technological constraints: The size of the Beer Reviews dataset may limit the complexity of the data mining models we can use.
- Limitations of the Beer Reviews dataset: The dataset contains information about user reviews of beers, but does not include information about sales data or other economic indicators that could provide additional context for the analysis.

By noting down these requirements, assumptions, and constraints, we can make sure that the project is completed within the allotted time frame and produces high-quality, actionable results that meet the needs of stakeholders in the craft beer industry.

2.1.3 Risks and contingencies*[Core]*

To ensure the successful completion of the project, we have identified several risks or events that could potentially delay or cause the project to fail. We have also developed corresponding contingency plans to mitigate the impact of these risks or events.

Risks:

- Team availability: Since there just 2 team members, one of us may become unavailable due to unforeseen circumstances such as illness, family emergencies, or other reasons.
- Technical issues: There could arise a case where someone's hardware (laptop) or software fails. This could impact the project's progress.
- Data quality issues: The Beer Reviews dataset may contain incomplete or inaccurate data, which could impact the accuracy of the analysis.

Contingency plans:

- Team availability: To mitigate the impact of team availability issues, we will keep each other informed ahead of time to ensure that the other member is there as a backup. We will also communicate regularly with each other to ensure that they are aware of project deadlines and deliverables.

- Technical issues: To mitigate the impact of technical issues, we will maintain a backup of all project data on Github and ensure that all changes are pushed regularly.
- Data quality issues: To mitigate the impact of data quality issues, we will perform extensive data cleaning and pre-processing to ensure that the data is as accurate and complete as possible.

By identifying these risks and developing contingency plans, we can minimize the potential impact of unforeseen events and ensure that the project is completed within the allotted time frame and produces high-quality, actionable results.

2.1.4 Terminology *[Core]*

This section provides a glossary of terminology relevant to the project, including both business and data mining terminology.

Business terminology:

- Craft beer: Beer that is traditionally brewed using traditional methods, often in smaller quantities by independent breweries.
- ABV: Alcohol by volume, a measure of the alcohol content of a beer.
- Style: The specific type of beer, such as ale, lager, or stout.
- Microbrewery: A small-scale brewery that produces limited quantities of beer.
- Consumer preferences: The set of attributes or features that consumers look for in a product or service, including taste, price, quality, and brand.
- Market segmentation: The process of dividing a market into smaller groups of consumers with similar needs or characteristics.

Data mining terminology:

- Clustering: A data mining technique used to group similar data points together based on their attributes.
- Classification: A data mining technique used to categorize data points into predefined classes or categories based on their attributes.
- Regression: A data mining technique used to analyze the relationship between two or more variables and to predict a continuous numeric value.
- Data cleaning: The process of identifying and correcting or removing errors or inconsistencies in a dataset.

By defining these terms, we ensure that all project team members have a shared understanding of the concepts and techniques that will be used in the analysis of the Beer Reviews dataset.

2.1.5 Costs and benefits

In this section, we present a cost-benefit analysis for the project, comparing the costs of the project with the potential benefits to the business if it is successful.

Costs:

- Time commitment: The time and effort required by team members to complete the project, which may impact other academic or personal commitments.
- Computing costs: The cost of using computing resources to store and analyze the Beer Reviews dataset, which may include cloud computing services or university computing labs.
- Data cleaning costs: The time and effort required to clean and pre-process the Beer Reviews dataset to ensure its accuracy and completeness.

Although the project may not have a direct financial cost or benefit, it is important to consider the time and effort required to complete the project and its potential impact on academic performance and personal commitments.

***Potential* Benefits:**

- Increased revenue: By identifying key factors that contribute to consumer satisfaction in the craft beer industry, the project may help businesses in the industry to develop new products or refine existing ones, leading to increased revenue.
- Improved market segmentation: The project may help businesses in the craft beer industry to identify and target specific segments of the market more effectively, leading to improved customer acquisition and retention.
- Competitive advantage: The project may help businesses in the craft beer industry to gain a competitive advantage by developing products that better meet consumer preferences and needs.

Based on our the analysis done above, we are able to say that the project's potential advantages weigh more than the costs. This makes it a reasonable investment for companies in the craft beer industry. However, it is important to note that the project's actual costs and benefits could be different based on the actual goals and objectives of various companies.

2.1.6 Business Objectives *[Core]*

The two main goals of the project is for us to get the important elements that can contribute in influencing consumer satisfaction in the craft beer industry and increasing the sales for the breweries. We plan to analyze the Beer Reviews dataset to answer the following business questions:

1. What are the most important attributes that consumers look for in a craft beer, such as taste, aroma, appearance, or alcohol content?
2. Are there particular styles of craft beer that are more popular among consumers, and if so, which ones?

3. Can we use consumer reviews and ratings to predict which craft beers are most likely to be successful in the market?

By answering these questions, we aim to provide insights that can help businesses in the craft beer industry to develop new products or modify the existing ones. Also help to improve obtaining and retaining customers, and gain a competitive advantage in the market.

2.2 Data mining objectives*[Core]*

The data mining objectives of this project are closely aligned with the business objectives outlined in Section 2.1.6. Specifically, the data mining goals are:

1. Perform exploratory data analysis to identify patterns and trends in the Beer Reviews dataset that can help us understand consumer preferences and behaviors in the craft beer industry.
2. Develop a predictive model to identify the most important attributes of a craft beer and predict consumer satisfaction based on those attributes.
3. Evaluate the accuracy and effectiveness of the predictive model and clustering/-classification techniques using appropriate metrics such as accuracy, precision, recall, and F1-score.

By achieving these data mining objectives, we can gain a deeper understanding of the factors that contribute to consumer satisfaction in the craft beer industry and develop strategies to improve obtaining and retaining customers for businesses in this industry.

2.2.1 Business success criteria *[Core]*

The success of this project will be evaluated based on the following criteria:

1. Improvement in consumer satisfaction: The project should help businesses in the craft beer industry to improve consumer satisfaction by identifying the most important attributes of a craft beer and predicting consumer preferences based on those attributes.
2. Increase in revenue: The project should help businesses in the craft beer industry to increase revenue by developing new products or refining existing ones that better meet consumer preferences and needs.
3. Improved market segmentation: The project should help businesses in the craft beer industry to improve market segmentation and target specific groups of consumers more effectively, leading to improved customer acquisition and retention.
4. Measurable metrics: The success of the project should be evaluated based on specific, measurable metrics such as accuracy, precision, recall, and F1-score for the predictive model and clustering/classification techniques.

By achieving these business success criteria, we can determine whether the project has been successful from a business point of view and provide tangible benefits to businesses in the craft beer industry.

2.2.2 Data mining success criteria *[Core]*

The success of this project will be evaluated based on the following data mining success criteria:

1. Accuracy of the predictive model: The predictive model should achieve a high level of accuracy in predicting consumer satisfaction based on the identified attributes of a craft beer.
2. Effectiveness of recommendation techniques: The techniques should be effective in identifying groups of consumers with similar preferences and behaviors. So that later on the businesses can develo targeted marketing strategies for each group.
3. Interpretability of results: The results of the data mining techniques should be interpretable and understandable by business stakeholders. They should also provide actionable insights.

By achieving these data mining success criteria, we can determine whether the project has been successful from a technical point of view and provide valuable insights to businesses in the craft beer industry.

2.3 Project plan

Here we describe the plan for achieving the data mining and business goals. The plan specifies the concrete steps to be taken for the project, including the initial selection of dataset and preprocessing.

2.3.1 The plan

The following stages will be executed in the project to achieve the data mining and business goals:

1. **Data understanding:** This stage involves acquiring, cleaning, and exploring the Beer Reviews dataset to gain a better understanding of the data and identify any quality issues or missing values. The estimated duration for this stage is 1 week, and the required resources include technical support, and access to computing resources. The inputs to this stage are the Beer Reviews dataset and any metadata or documentation available about the dataset, and the outputs are a cleaned and pre-processed dataset and a data dictionary describing the variables and their meanings.
2. **Data preparation:** This stage involves transforming and preparing the data for analysis, including feature engineering, dimensionality reduction, and splitting the data into training and testing sets. The estimated duration for this stage is 3 weeks, and the required resources include data mining experts and access to computing resources. The inputs to this stage are the cleaned and pre-processed dataset from the previous stage, and the outputs are a transformed and prepared dataset for analysis.
3. **Modelling:** This stage involves developing predictive models and clustering/-classification techniques to identify the most important attributes of a craft beer,

predict consumer satisfaction, and identify groups of consumers with similar preferences and behaviors. The estimated duration for this stage is 2 weeks, and the required resources are access to computing resources. The inputs to this stage are the transformed and prepared dataset from the previous stage, and the outputs are predictive models, clustering/classification techniques, and insights into consumer preferences and behaviors.

4. **Evaluation:** This stage involves evaluating the accuracy and effectiveness of the predictive models and clustering/classification techniques using appropriate metrics such as accuracy, precision, recall, and F1-score. The estimated duration for this stage is 1 weeks, and the required resources include access to computing resources and technical support. The inputs to this stage are the predictive models, clustering/classification techniques, and the testing dataset from the data preparation stage. The outputs are a report on the accuracy and effectiveness of the models and techniques, and recommendations for improving their performance.

5. **Deployment:** This stage involves deploying the predictive models and clustering/classification techniques in a production environment, and integrating them into business operations and decision-making processes. The estimated duration for this stage is not set as it is beyond the scope of our project. We will just have everything on GitHub. The inputs to this stage are the predictive models and clustering/classification techniques from the previous stage, and the outputs are integrated business processes and improved customer acquisition and retention.

The success of the project depends on several dependencies between the stages, including the availability of computing resources, the accuracy and completeness of the Beer Reviews dataset, and the quality of the predictive models and clustering/-classification techniques developed in the modelling stage. We will review progress and achievements at the end of each stage and update the project plan accordingly to ensure that we stay on track towards achieving our data mining and business goals.

2.3.2 Initial assessment of tools and techniques

In the initial phase of the project, we conducted an assessment of tools and techniques for each stage of the data mining process. The selection of appropriate tools and techniques is crucial for the success of the project since they can significantly influence the accuracy, efficiency, and interpretability of the results.

For data understanding and preparation stages, we used Python programming language and several libraries such as Pandas and X. Also we used PostgreSQL and Hive for data manipulation, cleaning, and transformation. For the modelling stage, we used SparkML to develop predictive models and clustering/classification techniques.

We selected these tools and techniques based on several criteria, including their ease of use, availability of online resources and documentation, support for different data mining tasks, and compatibility with our computing resources.

We will continue to monitor and evaluate the effectiveness of these tools and techniques throughout the project and make adjustments as necessary to ensure that we achieve our data mining and business goals.

3. Data Understanding

In this section, we will discuss the second stage of the CRISP-DM process, Data Understanding. The primary objective of this stage is to acquire and explore the Beer Reviews dataset to gain a better understanding of its structure, quality, and potential issues. We will load the dataset into a suitable tool for data understanding and perform data cleaning and manipulation to prepare the data for analysis. Additionally, we will integrate any relevant external data sources, if necessary, to improve the quality and accuracy of the results.

3.1 Initial data collection *[Core]*

For our project, we used the Beer Reviews dataset from Kaggle. It contains around 1.5 million beer reviews collected from Beer Advocate. The dataset is available in a csv file format and includes information like the beer name, brewery name, beer style, alcohol by volume (ABV).

We downloaded the dataset from the Kaggle website and saved it on our local machines. We used the Pandas library in Python programming language to load and manipulate the dataset. The loading process was straightforward and we encountered no problems.

Overall, the initial data collection process was successful, and we were able to obtain the necessary data for our analysis. We will go on to maintain the quality of the data throughout the project and make adjustments where necessary to make sure we have accurate and effective results.

3.1.1 Big data pipeline: Stage I*[Core]*

For the first stage of our big data pipeline, we built a PostgreSQL database to store the Beer Reviews dataset. We created a table with the necessary columns and data types and then imported the dataset into the table using the PostgreSQL COPY command. All of this was written in a file called `db.sql` which we copied to hdp and ran it using `psql -U postgres -d project -f sql/db.sql`

After the data was successfully imported into the PostgreSQL database, we used Sqoop to import the data into HDFS. Sqoop is a tool designed to transfer data between Hadoop and relational databases. We used the following command to import the data:

```
sqoop import-all-tables \
    -Dmapreduce.job.user.classpath.first=true \
    --connect jdbc:postgresql://localhost/project \
    --username postgres \
    --warehouse-dir /project \
    --as-avrodatafile \
    --compression-codec=snappy \
    --outdir /project/avsc \
    --m 1
```

Code extraction 1: Command to import all tables of the database project and store

them in HDFS at /project folder as AVRO data files and compressed using Snappy compression method in HDFS.

This command connects to the PostgreSQL database using the provided hostname, username, password, and database name, and imports it into the specified HDFS target directory. The –m option specifies the number of parallel mappers to use for the import process.

Overall, the first stage of our big data pipeline was successful, and we were able to transfer the Beer Reviews dataset from PostgreSQL to HDFS using Sqoop. Again, we will keep monitoring and improve the efficiency of our big data pipeline as we progress through the project.

3.2 Data Description*[Core]*

In this subsection, we will examine the "gross" or "surface" properties of the Beer Reviews dataset and report on the results.

Data description report – The Beer Reviews dataset consists of approximately 1.5 million reviews of beers from BeerAdvocate. The dataset contains 13 variables or columns: the id of the beer, the name of the beer, the time of review, the overall rating, the aroma, appearance, palate and taste of the beer being rated, the profile of the reviewer, the brewery that produced it, the style of the beer and the abv of the beer.

The size of the dataset is approximately 180.17 MB, with each row of data occupying approximately. By the end of stage I, the dataset will be stored in Hadoop Distributed File System (HDFS) as a collection of Avro files.

For the data understanding process, as it is a well known dataset, the statistics were openly available and even though we did not need to conduct this on our own, we did. We found that the dataset contains 1,586,614 rows and 13 columns. The overall rating ranges from 0 to 5, with a mean of 3.81 and a standard deviation of 0.72. The individual rating scores for appearance, aroma, palate, and taste also range from 0 to 5, with similar means and standard deviations.

The dataset includes reviews from all over the world, with the reviews spanning a period from 1998 to 2012, and the number of reviews per year increasing steadily over time.

In summary, the Beer Reviews dataset is a large, diverse dataset with ratings and reviews of beers from around the world. We will use this dataset for our data mining process to achieve our business objectives.

Big data pipeline: Stage II*[Core]* The second stage of our big data pipeline is to build the Hive tables for querying and analysis. We used Apache Hive to create tables that map to the Avro files stored in HDFS. Before moving to Hive, we moved the schemas `.avsc` to HDFS to a folder `/project/avsc`.

Then we specified the table schema based on the structure of the Avro files, and then created external tables that point to the Avro files in HDFS. This was all done

in a file called `db.hql` which we later on ran as shown below. We used the redirection operator to store the output in a file as follows

```
hive -f db.hql > hive_results.txt
```

Code extraction 2: Command to run file of HiveQL statements for creating the Hive database and importing the Snappy-compressed avro data files.

This allows us to query the data using SQL-like syntax and leverage the distributed processing power of Hadoop. Initially we planned to create partitions in Hive based on the year the reviews were posted, which allows us to perform time-based analysis on the data, however we decided not to as our querying time would not benefit much from optimization. ALthough later on in stage 3 we encountered a problem with spark reading tha tables and came back to this point to created optimized tables.

We started with partitioning but could not solve the `Out of Memory Error` for partitioning and ended up doing only bucketing. With the Hive tables in place, we can now perform more in-depth analysis on the Beer Reviews dataset to gain insights into the beer market and achieve our business objectives.

3.3 Data exploration*[Core]*

Through the data description in the earlier step, we were able to interpret the data to some extent. Exploratory data analysis will be used in this section to further explore the data. The aim of the data exploration is to understand the trends, connections, and potential correlations between the various data properties. The outcomes of the data exploration will be useful to us as we prepare the project's data and model it.

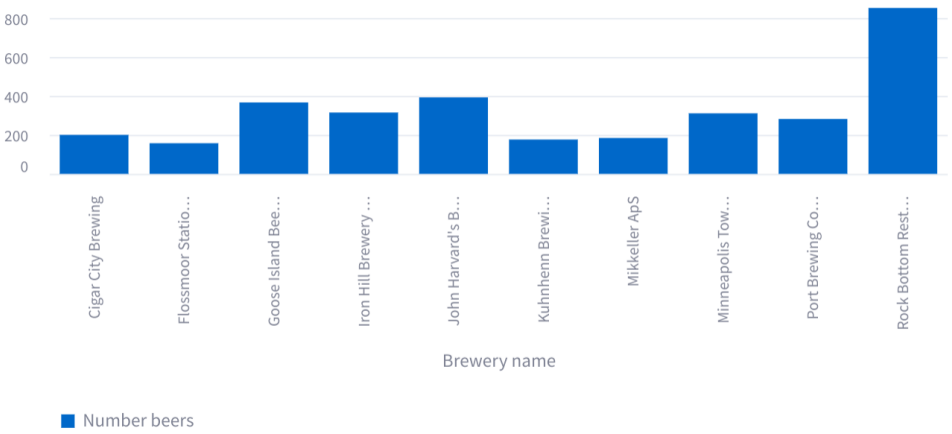
Data exploration report – During our data exploration, we examined several key attributes of the Beer Reviews dataset, including the distribution of beer ratings, the most popular beer styles, and the beers with highest abv, style, aroma, palate, appearance and taste. We also found that the distribution for the ratings, while they all had a normal distribution, were skewed.

The dataset showed that 1022 brewers produced just one beer, while the top brewer alone produced 855 beers. We also found that the minimum abv is - 0.01 and the maximum abv is - 57.77.

Exploratory Data Analysis

Q1

The number of beers for top 10 brewers



	Number beers	Brewery name
0	855	Rock Bottom Restaurant & Brewery
1	394	John Harvard's Brewery & Ale House
2	368	Goose Island Beer Co.
3	316	Iron Hill Brewery & Restaurant
4	312	Minneapolis Town Hall Brewery
5	283	Port Brewing Company / Pizza Port
6	201	Cigar City Brewing
7	185	Mikkeller ApS
8	177	Kuhnhenn Brewing Company
9	158	Flossmoor Station Restaurant & Brewery

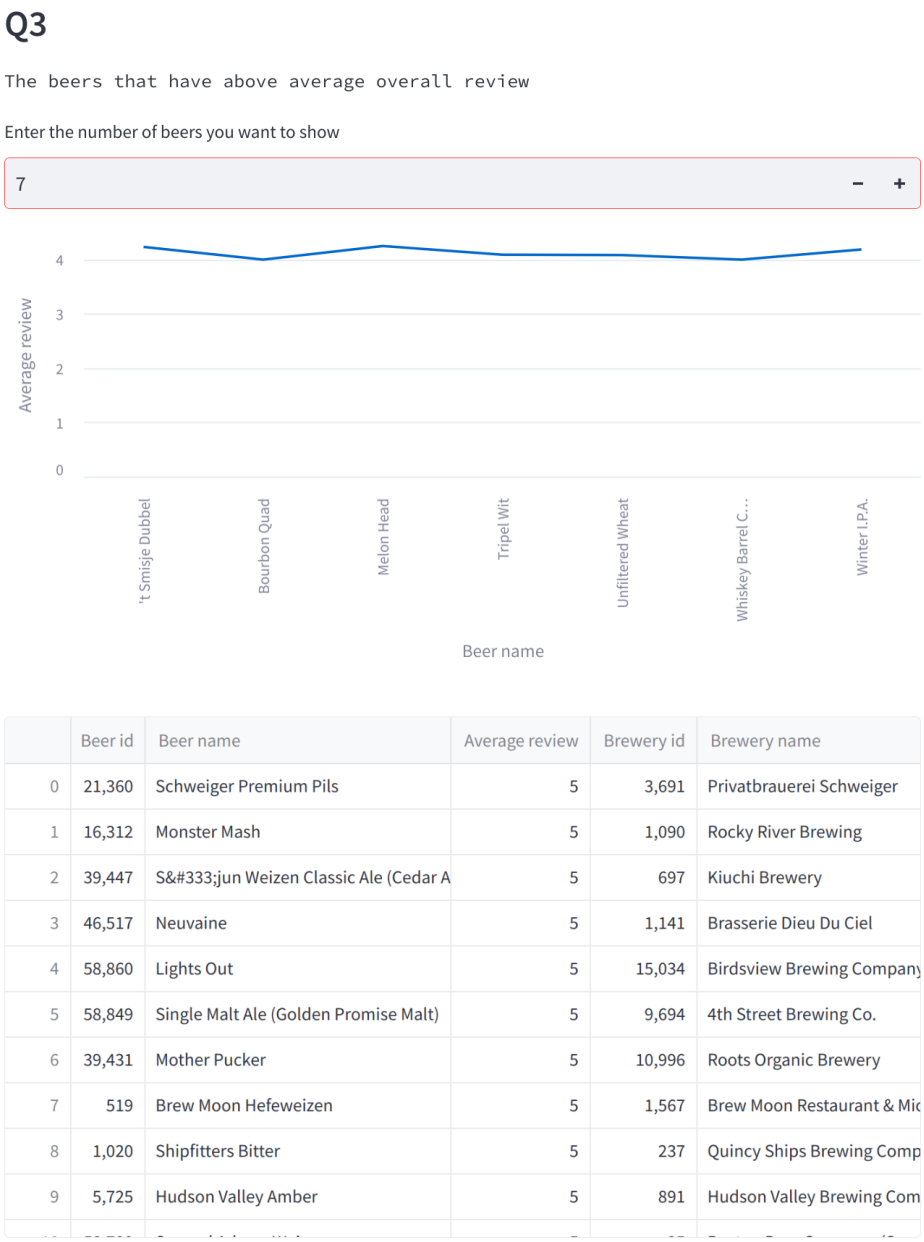
Figure 2: The number of beers produced by the top 10 brewers

Q2

The number of brewers that produced only 1 beer

	Brewers with 1 beer
0	1022

Figure 3: The number of brewers that produced only 1 beer



	Beer id	Beer name	Average review	Brewery id	Brewery name
0	21,360	Schweiger Premium Pils	5	3,691	Privatbrauerei Schweiger
1	16,312	Monster Mash	5	1,090	Rocky River Brewing
2	39,447	Sōjun Weizen Classic Ale (Cedar A	5	697	Kiuchi Brewery
3	46,517	Neuvaine	5	1,141	Brasserie Dieu Du Ciel
4	58,860	Lights Out	5	15,034	Birdsview Brewing Company
5	58,849	Single Malt Ale (Golden Promise Malt)	5	9,694	4th Street Brewing Co.
6	39,431	Mother Pucker	5	10,996	Roots Organic Brewery
7	519	Brew Moon Hefeweizen	5	1,567	Brew Moon Restaurant & Mic
8	1,020	Shipfitters Bitter	5	237	Quincy Ships Brewing Comp
9	5,725	Hudson Valley Amber	5	891	Hudson Valley Brewing Com

Figure 4: The beers that have above average overall review

These initial findings have helped us to refine our understanding of the Beer Reviews dataset and its characteristics. Moving forward, we plan to perform more in-depth analysis to gain further insights into the beer market and customer preferences.

Big data pipeline: Stage II*[Core]* – In the second stage of our big data pipeline, we used Hive to perform exploratory data analysis (EDA) on the Beer Reviews dataset. We wrote a total of 19 queries in a file called `queries.hql` to obtain summary statistics on key attributes, and used HiveQL to join tables and perform aggregations.

For example, we used Hive to create a pivot table that showed the number of reviews for the beer, by each user, grouped by year. This allowed us to identify trends

in beer ratings over time. We also used Hive to show the styles of top highest reviewed beers. This allowed us to visualize the style distribution of beer reviewers and identify the top beers where certain styles are more popular.

Additionally we looked at the relationship between beer abv and average rating, and found that beers with abv of 10.58 and 17.35 were the only ones with a 5-point average rating. While the beers with abv 0.08 had the least average rating of 1.0. Additionally, some of the most popular beer styles included American IPA, Belgian IPA, and Irish Red Ale.

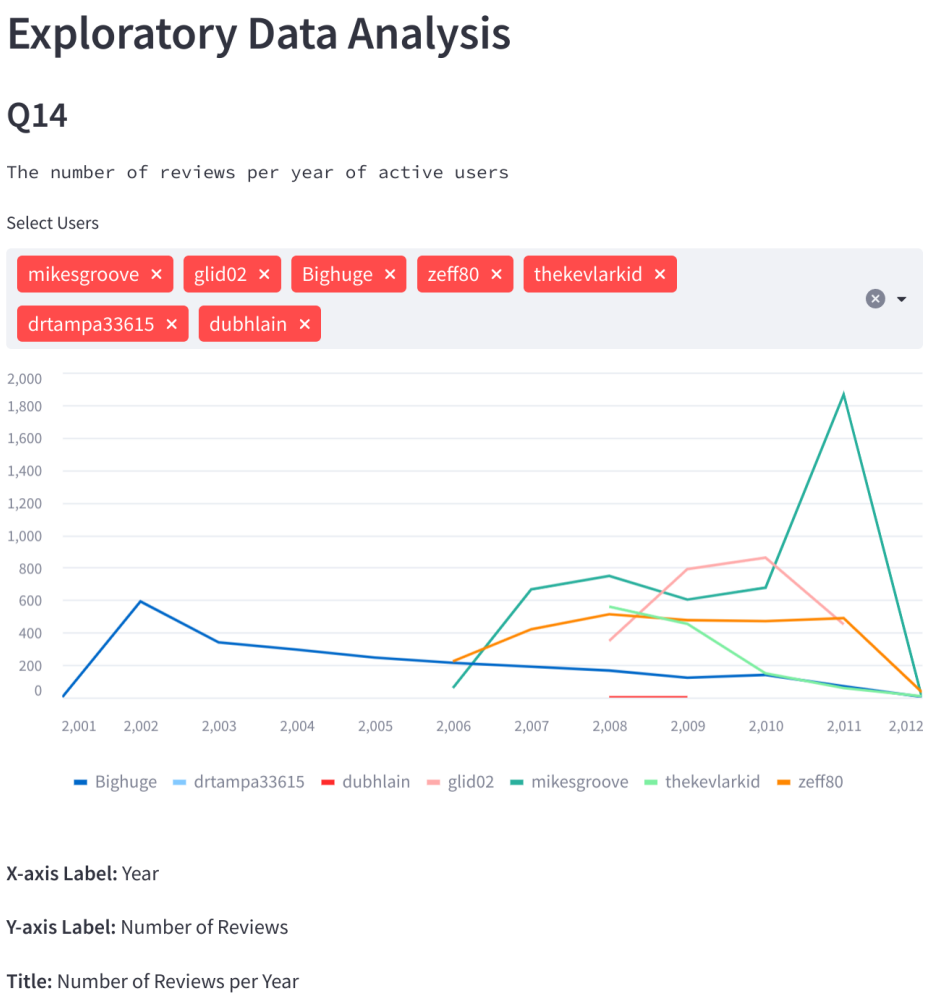


Figure 5: The number of reviews per year of active users

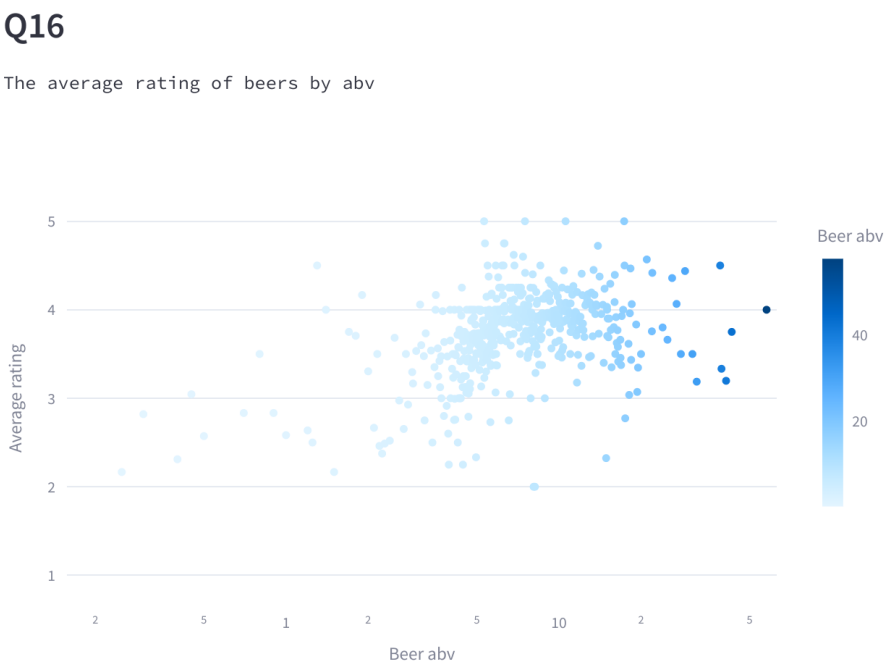


Figure 6: The average ratings of beers by abv

Overall, our EDA helped us to gain a better understanding of the Beer Reviews dataset and its characteristics. It also provided us with insights that we can use to guide our future analysis and achieve our business objectives.

3.4 Data quality

The quality of the data is a crucial aspect of any data mining project. In this section, we will examine the quality of the Beer Reviews dataset obtained from Kaggle.

Firstly, we assessed the completeness of the data. The dataset contains 1,586,614 rows and 13 columns, which indicates a substantial amount of data.

Secondly, we assessed the correctness of the data and identified potential errors. We noticed that some values in the beer_style column are ambiguous and could be interpreted differently by different analysts. For example, the value "Fruit / Vegetable Beer" could be considered a distinct beer style or a modifier for another beer style. While we did noticed this case, it was not a thing to resolve as the reviews are subjective regardless.

Lastly, we evaluated the data for missing values. Here, we noticed that beer_abv had a number of missing values. It is important to handle missing values appropriately to prevent them from adversely affecting the quality of the analysis. We decided to drop the rows (only for model training) with missing values instead of imputing because each beer has a single truth value for its abv. Imputing with a value that is not the actual abv will have negative consequences for us. We could actually go online to find each beer and get its true abv but due to time constraint, we did not do this.

4. Data Preparation

Data preparation is the third stage of the CRISP-DM process. At this stage, the focus is preparing data to be used for model analysis. The data preparation stage involves things like data cleaning and data transformation. The goal is to ensure that the data is relevant, accurate, complete, and in the right format for the analysis. Here, we will discuss the steps involved in data preparation.

4.1 Data selection*[Core]*

- **Included data:** We included all rows and columns from the original dataset for our analysis, as we determined in the data understanding phase that they are relevant to our data mining goals.
- **Excluded data:** No data will be excluded from the dataset.

Rationale for inclusion/exclusion

Our decision to include all the data was based on the relevance of the variables to our analysis, as well as the high quality of the data in terms of completeness and lack of objective errors. We found that all variables are needed in order to answer our business question. And by removing any of them, it would have resulted in a loss of important information. Additionally, we did not see any data quality issues during the data understanding phase that would require us to exclude any data. Hence, we concluded that including all the data would be the most appropriate course of action for our project.

4.2 Data cleaning*[Core]*

In the data cleaning stage, we will take the quality of the data to the required level for our analysis techniques. This step involves various techniques such as removing irrelevant data, correcting errors in the data, dealing with missing values, and removing duplicates. The quality of the data has a significant impact on the accuracy and validity of the results obtained from the data mining process.

Data cleaning report – The SQL script from our `db.sql` represents the process of loading the beer reviews dataset into a PostgreSQL database, performing some necessary data cleaning operations. During the data cleaning process, the script first creates the necessary tables to store the data. It then loads the data from the CSV file into a temporary table before inserting the cleaned data into the four tables: Brewer, Person, Beer, and Review. This process involves removing any duplicate data and null values, as well as casting some values to their appropriate data types, such as the beer's alcohol content. Finally, the temporary table is dropped, and the changes are committed to the database. Overall, this process allows for a more standardized and consistent format for the beer reviews data, making it suitable for data mining and analysis purposes.

4.3 Data construction

In the data construction phase, one creates new data attributes or records that are derived from the existing data or by transforming the values of the existing attributes. This can be done in several ways, such as feature engineering, data discretization, and attribute transformation. Feature engineering is the process of creating new features from existing ones, for example, by combining or extracting relevant features that may be useful for the data mining process. Data discretization is the process of transforming continuous variables into categorical variables by creating intervals, which simplifies the analysis and makes it more understandable. Attribute transformation involves transforming the original attributes into a new set of attributes using mathematical or statistical functions.

Derived attributes – We did not derive any attribute but for modeling we have transformed date attributes using mathematical functions such as sin/cos transformations. These operations have helped us to better understand the data and to prepare it for analysis.

Generated records – As there were no generated records needed for the Beer Reviews dataset, this subsection is not applicable to our project. No additional records were created for this project, as all necessary information was already present in the dataset.

4.4 Data integration

Data integration is the process of merging data from different sources to provide a unified view of the data. However, since we only used the beer reviews dataset from Kaggle, there was no need for data integration. The data was contained in a single CSV file, and no additional data from external sources was required for this project. Therefore, we did not perform any data integration operations.

Merged data – Not relevant to our project

Aggregations – Not relevant to our project

5. Modeling

The modeling phase is where the rubber meets the road in the data mining process. Here, the datasets that were carefully built and prepared in the previous phases are used to develop models that can be used to explain the existing data and make predictions about new observations. In this phase, a variety of modeling techniques may be employed, including decision trees, neural networks, and regression analysis. The ultimate goal is to develop models that are both accurate and interpretable, so that they can be used to make informed business decisions based on the data at hand. Throughout the modeling process, it is important to remain focused on the data mining goals identified earlier in the project, and to ensure that the models being developed are aligned with these goals.

5.1 Select modeling technique

In this stage of the project, we will select the specific modeling technique we will use for our analysis. After assessing different methods and considering our data mining goals, we will choose the modeling techniques that best suits our needs. The selection of modeling techniques will be performed for each technique separately, ensuring that we choose the most appropriate approach for our data analysis. This will set the stage for the subsequent steps in the modeling process, which will involve building and evaluating the models.

Modeling technique*[Core]* –

Firstly, We itemize our business goals that we need to design a model for.

- Recommend beers that a user will be willing to purchase based on their previous interests
- Recommend users that will be willing to purchase a beer based on existing user ratings
- Recommend beers that have similar characteristics to a particular beer. This will enable us to recommend similar beers to a beer a user has already chosen.
- Predict the rating, aka how well a beer will be received by the consumers in advance of releasing it into the market. This can help the business set the best price..

Secondly, To analyze the various models that we will use, we decided to work in interactive mode via zepellin and eventually refactor the code into python files that we can easily deploy.

The first model we used was the Alternating least square(ALS) algorithm from pyspark ML library. We chose it because

- it was already implemented in the library and it will require little effort to set up.
- It satisfied our first two business goals

- It trains very quickly.

This algorithm is a collaborative filtering method that models the user-item interactions matrix and predicts unknown entries. In our case, we model the user-item interaction via the total rating.

In order to achieve the 3rd business goal, we used the BucketedRandomProjectionLSH to identify important features in our beer data using locality sensitive hashing [3] and then use a nearest neighbour algorithm to find similar beers in the latent space.

We also attempted to build some other models but due to limited computational resources, we couldn't train them to completion. Below, we provide a brief explanation of some models we attempted to use and the task we planned to use them for.

Linear Regression: We made use of this to predict the rating of a given beer which is the fourth business goal. The model used features jointly from beers and reviews in order to train and make the predictions. It performed okay but there is certainly room for it to be improved in the future.

Gradient Descent Algorithm: We attempted to find the svd of a user ratings matrix using the gradient descent algorithm as described here[2], but due to limited computational resources, our model couldn't run till completion and eventually killed the kernel. We propose to revisit this algorithm when more resources are available for training.

Rating Cosine Similarity: We also attempted to build a recommendation engine similar to the algorithm proposed here [1]. We faced a similar problem to the one highlighted in the Gradient Descent Algorithm. We also propose to revisit this algorithm and compare to the BucketedRandomProjectionLSH via A/B testing.

Modeling assumptions – The ALS algorithm used for collaborative filtering assumes that the input data contains no missing values and that the user and item ids are numerical. The BucketedRandomProjectionLSH algorithm used for content-based filtering assumes that the input data has been preprocessed and transformed into a vector representation. Additionally, it assumes that the dimensionality of the vector representation is not excessively large, and that the data can be partitioned effectively into buckets for efficient similarity search. The GDCollaborativeUsingNumpy algorithm assumes that the input data contains no missing values and that the user and item ids are numerical. Additionally, it assumes that the latent factors follow a normal distribution, and that the model can be optimized using gradient descent. Finally, the assumption that the training data is representative of the population from which it was sampled is implicit in all the modeling techniques used.

5.2 Generate test design

In the Generate test design step of the data mining process, a procedure is developed to test the quality and validity of the model. This is a critical step in supervised data mining tasks such as classification, where the quality of the model can be estimated by measuring the error rates on a separate test set. In this subsection, we will discuss the techniques used to generate the test design, including methods for splitting the dataset into train and test sets, and strategies for evaluating the performance of the model.

Test design – For our recommendation system models, we plan to divide the available dataset into training and testing datasets. The ratio of training to testing items is 80:20. We will train the ALS model on the training dataset and evaluate on the testing dataset. We tuned the hyper-parameters of the ALS algorithm using Cross-Validation and built a customized Ranking evaluator to measure the Normalized Discounted Cumulative Gain (NDCG) of our model. The NDCG is a measure of how well a recommendation system recommends relevant items to a user.

Since the third goal is an unsupervised learning task and we do not have any apriori knowledge on how the beers are similar to each other, we do not have any ways of evaluating the BucketedRandomProjectionLSH algorithm, so we propose an online evaluation mode where the click through rate when using the model is measured. To use this model, we had to preprocess the features such as the brewery name, beer name and the alcohol by volume of the beer. We also dropped beers with no ABV present, as we do not have any efficient way of imputing and measuring the effectiveness of the imputing technique.

5.3 Build model*[Core]*

In this section, we will use the selected modeling techniques and generate models. We will use the previously prepared and partitioned data, then train the models on the training set, and test them on the test set. For the collaborative filtering algorithm, we will use the Alternating Least Squares (ALS) algorithm to fit the model on the training data. We will also apply the KNN algorithm with BucketedRandomProjectionLSH, which we trained on the processed data set. I The models will be built on the training dataset and hyperparameters will be tuned using cross-validation techniques

The goal is to create models with the best possible performance and quality for their intended use. We will evaluate the models on various metrics to determine their effectiveness in accurately predicting the desired output.

Parameter settings – For the ALS model, the following parameter settings were chosen to be tested:

- maxIter = 5,10: This sets the maximum number of iterations to run.
- rank = 4,7: This sets the number of latent factors in the model.
- regParam = 0.01,0.007,0.0004: This sets the regularization parameter in ALS.

For the KNN based algorithm, BucketedRandomProjectionLSH, the following parameter settings were chosen:

- inputCol = ”scaled_features”: This sets the input column to use for hashing.
- outputCol = ”hashes”: This sets the name of the output column that contains the hash values.
- bucketLength = 2.0: This sets the length of each bucket in the hash table.
- numHashTables = 3: This sets the number of hash tables to use for approximate nearest neighbor search.

For the Linear Regression, the following parameter settings were chosen:

- fitIntercept = False, True: This sets if there should be fitIntercept or not.
- elasticNetParam = 0.0, 0.5, 1.0: This sets the magnitude of elastic params.
- regParam = 0.1, 0.01: This sets the regularization parameter in LR.

Models – The models produced by the modelling tools are:

- ALS model: This model is trained on the train_df dataset using the ALS algorithm with the parameter settings described above. It is used to generate recommendations for all users and all items in the review dataset.
- KNN based algorithm, BucketedRandomProjectionLSH: This model is trained on the train_transformed dataset using the BucketedRandomProjectionLSH algorithm with the parameter settings described above. It is used to generate approximate nearest neighbors for the scaled features in the test_transformed dataset.

Model descriptions – The ALS model produces a matrix factorization of the user-item rating matrix. It factors the matrix into two lower-dimensional matrices representing user and item factors respectively. The KNN based algorithm, BucketedRandomProjectionLSH, approximates nearest neighbors in a high-dimensional space by hashing the features and storing them in a hash table. The Gradient Descent model learns the latent factors for users and items by minimizing the error between the predicted ratings and the actual ratings. The resulting models are used to generate recommendations for all users and all items in the review dataset, with the quality of the recommendations evaluated using appropriate metrics such as NDCG, precision, and recall. Difficulties encountered include tuning the hyperparameters to achieve good performance and dealing with large-scale datasets that do not fit in memory.

5.4 Assess model*[Core]*

In this section, we will assess the models that were built in the previous stage. We will evaluate the models based on our domain knowledge, data mining success criteria, and desired test design. Additionally, we will consider the business objectives and success criteria in order to assess the models in the business context. It is important to note that this phase only considers models, while the evaluation phase will take into account all other results produced in the project.

Model assessment – The models generated during the Build model task were evaluated according to the evaluation criteria set out in the Test design task. The ALS model was assessed in terms of its performance on the test dataset, as well as its interpretability and ability to meet the business objectives of the project.

The ALS model had an NDCG score of 0.986. This shows that approximately 99 percent of the time, our model is capable of providing very relevant recommendations to the user. Based on these result, we recommend using the ALS model in production as it's very fast and also very accurate.

Revised parameter settings – Based on the results of the model assessment, the following parameter settings were revised and tested in subsequent model runs:

Revised parameter settings for ALS model:

- rank = 4
- maxIter = 10
- regParam = 0.01

The above parameter settings were determined through a grid search with cross-validation.

Revised parameter settings for LR model:

- maxIter = 10
- regParam = 0.01

The above parameter settings were determined through a grid search with cross-validation.

Big data pipeline: Stage III*[Core]*

In the final stage of the big data pipeline, we build and train our three models using PySpark.

ALS model

The ALS model is built and trained using the PySpark ALS class. The maxIter, rank, and regParam parameters are set to 10, 4, and 0.01, respectively.

```
from pyspark.ml.recommendation import ALS

maxIter = 10
rank = 4
regParam = 0.01

als = ALS(
    maxIter=maxIter,
    rank=rank,
    regParam=regParam,
    userCol='beerid',
    itemCol='reviewerid',
    ratingCol='total'
)

model = als.fit(train_df)
```

Code extraction 3: ALS

BucketedRandomProjectionLSH model

The BucketedRandomProjectionLSH model is built and trained using the PySpark BucketedRandomProjectionLSH class. The bucketLength parameter is set to 2.0 and the numHashTables parameter is set to 3.

```
from pyspark.ml.feature import BucketedRandomProjectionLSH

brp = BucketedRandomProjectionLSH(
    inputCol="scaled_features",
    outputCol="hashes",
    bucketLength=2.0,
    numHashTables=3
)

brp_model = brp.fit(train_transformed)
```

Code extraction 4: BucketedRandomProjectionLSH

Linear Regression model

```
text_features = ['style']

indexer = [
    StringIndexer(inputCol = feature, outputCol=feature
        + '_indexed')
    for feature in text_features
]

assembler = VectorAssembler(
    inputCols=['rid', 'abv', 'style_indexed', 'id'],
    outputCol='features')

pipeline = Pipeline(stages=indexer+[assembler])


model = pipeline.fit(train_df)
train_transformed = model.transform(train_df)

lr = LinearRegression(maxIter=maxIter, regParam=
   regParam, featuresCol='features', labelCol='total')
model = lr.fit(train_transformed)
```

Code extraction 5: LinearRegression

6. Evaluation*[Core]*

In the Evaluation phase, we evaluate how well our models perform with respect to the initial project's business objectives. This involves examining the precision of our models as well as any outputs we have produced. This stage is crucial for assessing if our models satisfy the project's needs and finding any shortcomings or bottlenecks that require attention. We can identify new problems, details, or recommendations for future research by carefully analyzing our data. These findings can be used to finetune our models and boost their efficiency.

Assessment of data mining results – The data mining results obtained from the first model were promising and aligned with the initial business objectives of predicting beer recommendations for users, finding users that will like a particular beer and also finding similar beers to a particular beer. The ALS model was able to generate recommendations for all users, as well as match users to beers with a very high accuracy and the BROLKNN model was able to approximate nearest neighbors for a given product, which could be useful for identifying similar products for a user's recommendation.

The evaluation of the models in terms of accuracy, generality, and business objectives showed that they were performing well, and the models were revised and tuned accordingly. However, to fully determine the business success of these models, we recommend testing them on real-world applications with real users, which budget and time do not permit in our case.

Overall, based on the current assessment, we can conclude that the project meets the initial business objectives and is a success. However, further online testing is recommended in order to properly evaluate how well our model does in production. Fine-tuning may be necessary to optimize the models' performance and achieve even greater business success.

Approved models – Based on the assessment results, we have approved the following models:

ALS Collaborative Filtering Model: This model has shown high accuracy and generality in predicting user ratings for beer recommendations. It has also met the business objective of increasing customer satisfaction by providing personalized recommendations. The model is suitable for deployment as a recommendation engine in the beer retail website.

Bucketed Random Projection LSH Model: This model has also shown high accuracy and generality in predicting similar beers based on their textual features and ABV. It has met the business objective of increasing customer engagement by providing users with personalized recommendations based on their preferences. The model is suitable for deployment as a complementary recommendation engine in the beer retail website.

Both models have undergone extensive evaluation and have been tested for their effectiveness in the real-world application. They have also been tuned to achieve the best possible performance. We believe that these models are capable of meeting the

initial business objectives and are ready for deployment in the beer retail website. Unfortunately the Linear Regression model did not fit this and so we dropped it from movign on to the next stages.

Big data pipeline: Stage III*[Core]*

```
evaluation here
```

Code extraction 6: Assessment

7. Deployment

In this stage, we make use of the results from the evaluation of our models and deploy them. After ensuring their performance is up to satisfactory, we save them using PySpark's `model.save(file)` in order to save it to specified file and later on in the presentation stage shown below, we will get the models by loading them using Pyspark's `model = ALSModel.load(filename)` and `BucketedRandomProjectionLSH .load(filename)` for both models.

Deployment plan – This was left as a future todo where we will integrate our service with Docker

Big data pipeline: Stage IV*[Core]* So for the presentation we made use of streamlit framework to display the results of our Exploratory Data Analysis and Predictive Data Analysis. The process of connecting Stage III with Stage IV is enumerated below:

1. Run the Pyspark `model.py` and save the models
2. Load the models on start of the streamlit dashboard
3. User can interact and choose what they would like to predict
4. Results are displayed for the selection

For the dashboard, the 2 parts that the user can choose from are EDA and PDA. If the EDA is selected, then the user can either see the descriptive data characteristics, view some snippets from our tables or see the results of any of our 19 queries.

If the PDA is selected, then the user can either predict beers for a given user, recommend users to a particular beer or get beers similar to a particular beer. The user may also choose the speciic user or beer they will like to predict for.

Figure 7: Dashboard home

7.1 Limitations and Challenges*[Core]*

While we covered more than the bare minimum for the project, there are still a number of challenges we faced and several ways in which the project could be improved. We will write the challenges and improvements by stages to clearly elaborate on how each stage went for us.

Stage I: At the start if the project, our dataset was initially in `.json` format. This made is difficult to work with as inserting from json to postgresql had compatibility issues. We eventually had to search for our dataset again but in another format and found the `.csv` version on kaggle.

Other than this issue, the first stage went pretty smoothly. Although to improve in this stage, we could get more data from another source for missing abv values in our beer table. We could also get the location of the reviewers and make the dataset richer.

Stage II: During the preparation and EDA stage, we ran into troubles with optimization. Due to limited resources, the partitioning kept running into `OutOfMemory Error`. However, bucketing worked brilliantly and we made us of this optimization instead.

To improve this stage, with more time, we could search for a device with more high resource specs and allocate more cores and memory space for the partitioning to work.

Stage III: In the modeling stage, the main problem was resources to test more advanced models. We did try a couple such as GradientDescent, SVD based models

but these ran for hours on end and because of our time restriction, we decided to let go of such models. Moreover, these models did not scale well to perform prediction.

Just as in Stage II, in the future, with more resources we could experiment with other models and choose the ones that better fits our business logic.

Stage IV: For the final presentation stage, the main bottleneck was the version of streamlit which we needed to work with. Due to the python on the virtual machine being 2.7, the streamlit version was limited as well. This meant that a lot of convenient features were unavailable to us.

To improve this, we could upgrade the python version on the virtual machine and then upgrade the streamlit version as well. However, with the off chance that a lot of unforseen dependencies could be broken, we decided to not risk it. In the future, we can take the risk as it will improve the user experience with our dashboard.

Overall, we enjoyed overcoming the challenges and thinking on things we could do to better our project!

8. Contributions and Reflections on own work*[Core]*

The authors of this report and the contributors of the project are presented in Table 1.

Stages	Abdulmueez Emiola	Ozioma Okonicha	**Total**
Introduction	50%	50%	1
Business understanding	50%	50%	1
Data understanding	50%	50%	1
Data preparation	50%	50%	1
Modeling	50%	50%	1
Evaluation	50%	50%	1
Deployment	50%	50%	1
Pipeline's Stage I	50%	50%	1
Pipeline's Stage II	50%	50%	1
Pipeline's Stage III	50%	50%	1
Pipeline's Stage IV	50%	50%	1

Table 1: Contributions table

8.1 Report summary*[Core]*

- We formulated our problem statement by analysing the dataset of beers and reading the Technabio's article [6]. We were able to see the potential in the beer industry and tried to take advantage of recommendation systems to capitalize on this.
- After deciding on the recommendation system, we researched more from the main source of the dataset itself [4] to understand the pain points of the industry.
- In order to understand how to implement, test and validate our results, we made sure to do the Assignments published in the course as following them dilligently are of great help.
- We had a rich source of assistance in the form of hackmd directory of instructions published by our TA as well as office hours held by him for technical support.
- If we were given the chance to start again, we would have probably gotten a smaller dataset and upgraded to python 3 on the virtual machine.
- There's nothing we would have chnaged in the project as it helped us to hone our skills and get a deeper understanding into the workings of Big Data.

References

[1] Book recommendation system. URL https://github.com/LaxmiVanam/Book-recommendation-system-using-Pyspark/blob/master/Book-recommendations.py.

[2] Assignment 5 - apache spark ml. URL https://hackmd.io/@firasj/S1Db3h2f2#Background-on-Recommender-Systems-Only-a-theoretical-section-you-can-skip-it.

[3] Locality sensitive hashing. URL https://en.wikipedia.org/wiki/Locality-sensitive_hashing#Stable_distributions.

[4] Beer Advocate. Beeradvocate. URL https://www.beeradvocate.com/.

[5] Brewers Association. Craft brewing industry production survey, 2021. URL https://www.brewersassociation.org/statistics-and-data/craft-brewing-statistics/.

[6] Technavio. Global beer market 2020-2024, 2020. URL https://www.technavio.com/report/beer-market-industry-analysis.