Brewery Operations and Market Analysis

Vincenzo Rocchi - 664957

March 2024 - DDAM exam

1 Dataset Introduction

The chosen dataset is taken from kaggle (Click here to visit the dataset on Kaggle and is a comprehensive analysis of Brewing Parameters, Sales Trends, and Quality Metrics in Craft Beer Production with data ranging from 2020 to 2024. It includes 8 different selling locations for the city of Bangalore and it's structured as follows:

- Brewing Parameters: Includes crucial brewing factors such as fermentation time, temperature, pH level, gravity, and ingredient ratios. These parameters are pivotal in understanding the brewing process and its impact on the final product.
- Beer Styles and Packaging: The dataset categorizes beers into various styles like IPA, Stout, Lager, etc., and records the type of packaging used (kegs, bottles, cans, pints).
- Quality Scores: Each batch is rated for its quality on a scale, offering insights into the success and consistency of different brewing approaches.
- Sales Data(USD): Detailed records of sales figures, providing a window into the market performance of different beer types across various locations in Bangalore.
- Supply Chain and Efficiency Metrics: Tracks aspects like volume produced, total sales, brewhouse efficiency, and losses at different stages (brewing, fermentation, bottling/kegging), crucial for supply chain analysis and operational optimization.

2 Dataset Preparation

To optimize data processing efficiency, Apache Arrow was enabled for data transfers between Python and the Spark execution engine. The dataset was loaded in the columnar Parquet format, and the data is represented as a PySpark DataFrame for subsequent preparation and analysis.

Table 1: Features of the Brewery Dataset

Feature Name	**Data Type**	**Description**
Batch_ID	Unique Identifier	Unique identifier for each beer batch
Brew_Date	Date	Date of brewing
Beer_Style	Categorical	Style of beer (IPA, Stout, etc.)
SKU	Categorical	Packaging type (Kegs, Bottles, etc.)
Location	Categorical	Geographical location of sale
Ferm. Time	Numerical	Fermentation duration (days)
Temperature	Numerical	Average brewing temperature (Celsius)
pH_Level	Numerical	Acidity/alkalinity level
Gravity	Numerical	Density compared to water
Alcohol Cont.	Numerical	Alcohol percentage
Bitterness	Numerical	IBU (International Bitterness Units)
Color	Numerical	SRM (Standard Reference Method)
Ingredient Ratio	Numerical	Ratio of malt, hops, etc.
Volume Produced	Numerical	Volume of beer in the batch (liters)
Total Sales	Numerical (Currency)	Sales generated
Quality Score	Numerical (Rating)	Overall quality (out of 10)
Brewhouse Efficiency	Numerical (Percentage)	Brewing process efficiency
Loss During Brewing	Numerical (Percentage)	Volume loss during brewing
Loss During Fermentation	Numerical (Percentage)	Volume loss during fermentation
Loss During Bottling/Kegging	Numerical (Percentage)	Volume loss during bottling/kegging

2.1 Feature Engineering and Transformation

The original "Ingredient_Ratio" column was split into separate numerical columns ("water", "grains", and "hops") to facilitate ingredient analysis. These extracted values were cast as double-precision floating-point numbers for computational compatibility. Additionally, two derived metrics were calculated: "USD_per_Liter" offers insights into production cost efficiency, while "Brewing_efficiency" provides a measure of process effectiveness. These transformations enhance the analytical capabilities of the dataset, potentially supporting cost and process optimization aspects of the data mining task.

2.2 Feature Encoding

The 'fermentation_time' feature was discretized into bins such as "Short", "Medium", "Long" and then encoded to enhance its usability in data analysis. Additionally, categorical features like 'Beer_Style', 'SKU', and 'Location' were one-hot encoded, resulting in new binary features representing each unique category within those columns.

3 Exploratory Analysis and Clustering

Initial exploration of the dataset revealed limited direct linear relationships among the features. Further analysis of feature distributions suggests the potential for identifying patterns that may not be immediately apparent through simple correlation metrics. Clustering techniques will be employed to uncover potential groupings within the data based on shared characteristics across multiple dimensions.

3.1 Clustering

Two distinct sets of features were explored for their suitability in clustering:

Feature Set	**Features**
Production Factors	Fermentation_Time, Temperature, pH_Level, grains, hops, Brewing_efficency
Product Characteristics	Alcohol_Content, Bitterness, Color

3.1.1 Approach

To prepare the data for the K-Means clustering algorithm, a preprocessing pipeline was constructed. This pipeline involved combining individual features into a vector representation using a VectorAssembler, followed by feature standardization with a StandardScaler. The elbow method at figure 1a and figure 2a was employed to guide the selection of the appropriate number of clusters (k), and cross-validation (CrossValidator) was used to refine the clustering model's

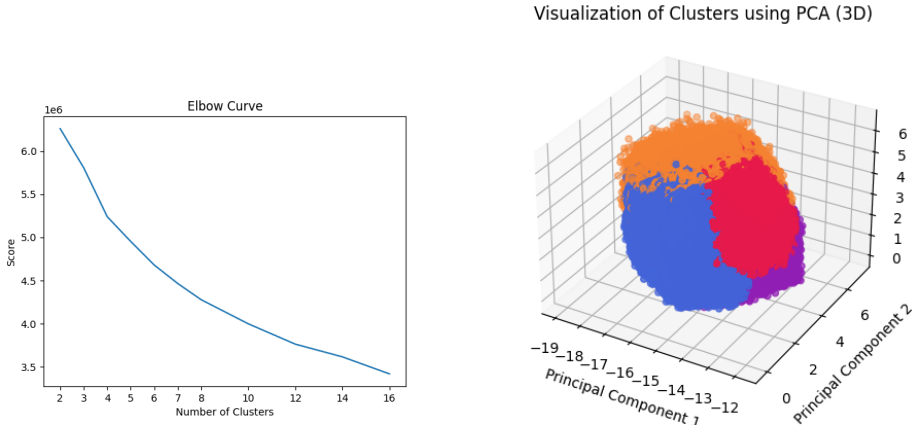
hyperparameters, including the initialization mode. Initial analysis points towards the presence of potentially meaningful clusters within the data.

3.1.2 Results and Analysis - Production

Cross-validation indicated that the optimal number of clusters is 4, based on the analysis of the silhouette metric. The K-Means initialization mode was reported as k-means++. A silhouette score of 0.216 suggests a moderate degree of cluster cohesion, indicating the presence of some structure within the dataset. The cluster distribution table reveals a noticeable imbalance in cluster sizes, with Cluster 1 being the largest, followed by clusters 3 and 0. Cluster 2 is significantly smaller.

3.1.3 Results and Analysis - Product

Cross-validation determined that 8 clusters provide the best balance for the product feature data based on the silhouette score. The K-Means initialization mode was reported as k-means++. A silhouette score of 0.414 indicates a moderate cluster cohesion, suggesting detectable structure within product features, though not an extremely strong separation. The cluster distribution table reveals a noticeable imbalance in cluster sizes, with Cluster 5 and 6 being the smallest followed by Clusters 4 and 7, while the others show an approximately identical row count.



(a) KMeans production batch (b) KMeans production batch

Figure 1: Comparison of KMeans production batches

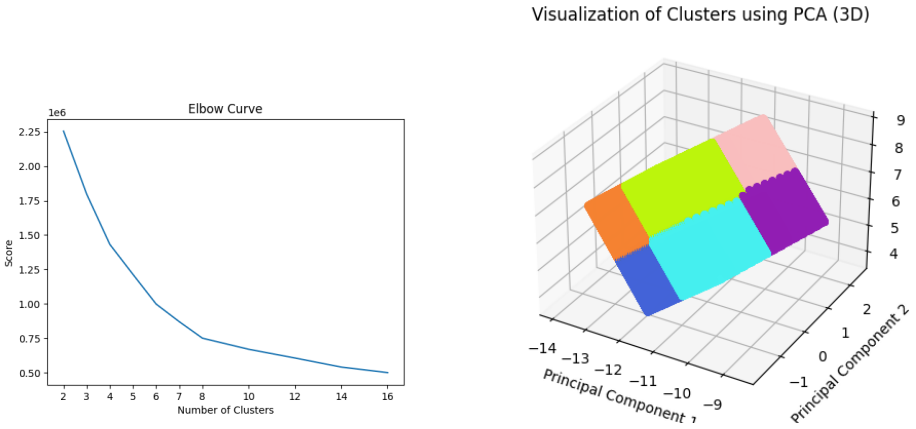
3.1.4 Visualization with Principal Component Analysis (PCA)

PCA was applied to reduce the dimensionality of the features used in clustering to three dimensions, enabling a 3D visualization as in figure 1b and in figure 2b. This technique aims to provide a visual representation of potential separation

between clusters. The fitted PCA model transformed the dataset, and a 3D scatterplot was created with each data point's coordinates derived from the first three principal components. Cluster assignments were color-coded to help identify any visual groupings. Additionally, the explainedVariance attribute of the PCA model was examined to understand the relative importance of each principal component.

For the **production features** the component weights for the top three principal components are very close in value (0.1669, 0.1668, 0.1667). This indicates that the original features contribute relatively evenly to the variation captured within these components. The lack of a single dominant component suggests Multidimensional Relationships or Limited Linear Separability

For the **product features** the component weights are remarkably similar across the top three components (0.3338, 0.3333, 0.3328). Suggesting the same assumptions as for the production part.



(a) KMeans product batch

(b) KMeans product batch

Figure 2: Comparison of KMeans product batches

3.1.5 Centroid Comparison Analysis - production

Clusters 1 and 3 exhibit the longest average fermentation times, with cluster 2 having the shortest. Cluster 0 favors the lowest average fermentation temperature, while cluster 2 favors the highest. Clusters 1 and 3 fall in between these extremes. The average pH level shows only slight variations across clusters, suggesting this feature might be less influential in the cluster differentiation. Gravity, alcohol content, bitterness, and color follow a similar pattern across clusters, indicating potential correlations between these features. Cluster 1 has the highest average 'Volume Produced,' closely followed by clusters 3 and 0. Cluster 2 has a significantly smaller volume on average. Noteworthy differences exist in 'Brewhouse_Efficiency' with cluster 2 having the highest average, and cluster 0 the lowest.

3.1.6 Centroid Comparison Analysis - product

A centroid analysis was performed to identify the defining characteristics of the eight clusters identified within the product feature data. The mean values of each feature for each cluster revealed several noteworthy trends. Clusters 1, 6, 3, and 5 exhibited longer average fermentation times (approximately 14.5 days) compared to clusters 2, 4, 7, and 0. Similarly, clusters 1, 3, 6, and 7 demonstrated a preference for slightly warmer average fermentation temperatures (approximately 20°C). Across all clusters, pH levels remained relatively consistent, suggesting minimal influence of this factor in cluster differentiation.

Significant variations emerged in alcohol content, bitterness, and color. Cluster 6 displayed the highest average alcohol content, followed by cluster 5, while clusters 1, 2, and 4 exhibited the lowest levels. Bitterness and color trends were correlated, suggesting a potential relationship between these two features. Analysis of production volumes revealed substantial differences, with clusters 1, 6, 3, and 5 demonstrating higher averages than the remaining clusters. Brewhouse efficiency did not show major differentiation across the clusters. Surprisingly, the ingredient ratios (water, grains, hops) remained relatively similar across clusters

3.2 Feature exploration

3.2.1 Correlation - pre clustering

In this analysis, PySpark's corr() function was employed to compute correlation matrices for the relevant dataset. To comprehensively assess the relationships between numerical features, both Pearson's and Spearman's correlation methods were utilized. Pearson's correlation coefficient quantifies the degree of linear association between variables, while Spearman's correlation coefficient evaluates the strength of a monotonic relationship, which may be linear or nonlinear.

Despite considering the potential for non-linear relationships, both correlation matrices exhibited negligible correlation among features. A more comprehensive exploration of these features will be conducted in the Feature Analysis chapter , as for now the results can be assessed in figure 3, only regarding the Pearson's coefficent.

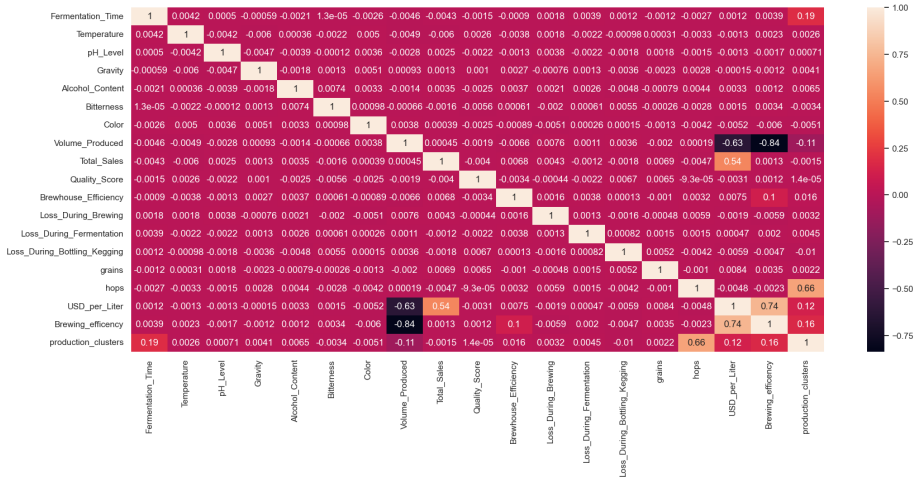


Figure 3: Pearson's correlation matrix - production clusters

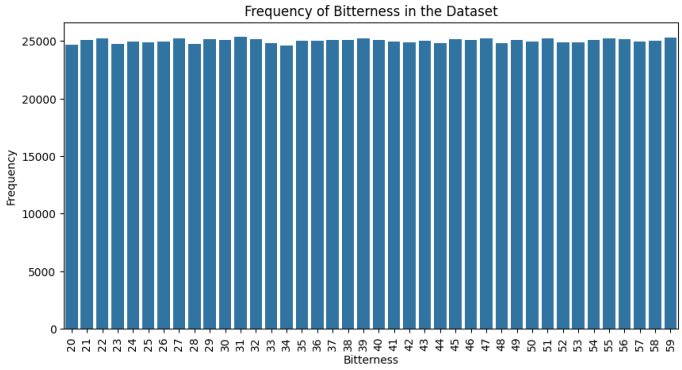
3.2.2 Correlation - post clustering

As detailed in Section 3.1.1, incorporating the cluster labels derived for production and product batches was anticipated to yield more informative correlation matrices. However, the influence of the low silhouette score, indicative of potentially weak cluster separation, is evident in the resulting correlation matrices. While the majority of features exhibit negligible correlation, the hops ratio feature stands out as exhibiting a significant correlation. Interestingly, fermentation time and brewing efficiency also show slight correlations with the production clusters. In the product domain, alcohol content, bitterness, and color exhibit slight negative correlations.

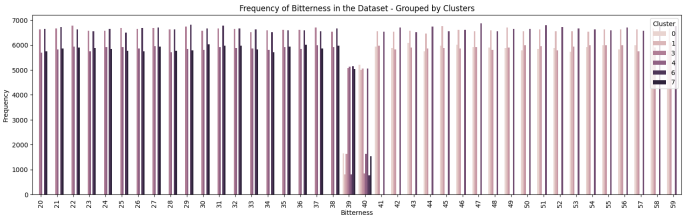
3.2.3 Feature analysis - low-distinct continous

Analyzing the distribution of features is essential for understanding the characteristics of a dataset. When data exhibits natural groupings or is divided into clusters, exploring feature distributions within these clusters can yield far greater insights. By examining how a feature's distribution varies across different clusters, we can potentially uncover patterns that differentiate clusters, identify anomalies or unexpected trends within specific subsets of the data, or gain insights into differential relationships with target variables in a supervised learning context.

We'll focus, in this part, on the 'Bitterness', "Grains and Hops" and "Fermentation Time" features. We will first examine its overall distribution in the dataset and subsequently investigate its distribution within the context of both 'production_clusters' and 'product_clusters'. This strategy aims to determine whether the distribution of those features varies significantly across different production or product groups.



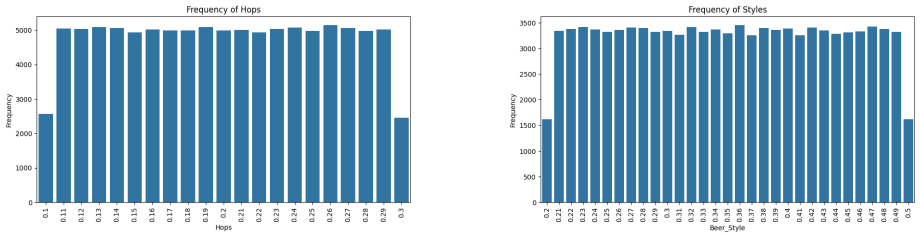
(a) Bitterness



(b) Bitterness product

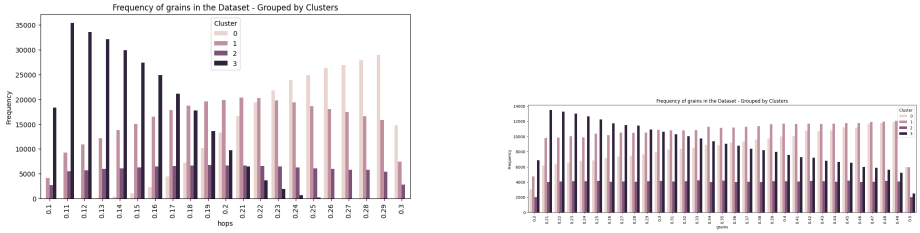
Figure 4: Bitterness comparison

The overall distribution of the **'Bitterness'** (4) feature initially appears surprisingly uniform. However, upon incorporating the 'product_clusters' information, a fascinating pattern emerges. We observe a clear breakpoint at a Bitterness level of 39. Clusters 2 and 5 co-occur with other clusters at all bitterness levels but are strikingly absent above the 39 threshold. Conversely, clusters 0 and 1 are exclusively found at bitterness levels exceeding 39. Interestingly, clusters 6 and 3 span the entire 'Bitterness' range, while cluster 7 seems confined to values below 40.



(a) Hops (b) Grains

Figure 5: Grains and Hops comparison



(a) Hops production (b) Grains production

Figure 6: Grains and Hops production comparison

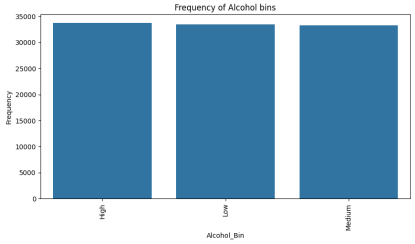
Similar to the 'Bitterness' feature, the **'Grains'** and **'Hops'** features (fig 5 and fig. 6) exhibit relatively flat distributions when analyzed in isolation. However, when we consider these distributions within the context of 'production_clusters', compelling patterns and trends emerge. Clusters 1 and 2 display distributions resembling a normal distribution, although cluster 1 shows a slightly increasing trend. Strikingly, cluster 3 exhibits a strictly decreasing pattern for both 'Grains' and 'Hops', while cluster 0 exhibits a sharply increasing trend.

The distribution of the **'fermentation_time'** feature, both in its original continuous form and as a binarized representation, displayed a relatively flat pattern. No discernible trends or patterns were observed when analyzed in the context of either product or production clusters.

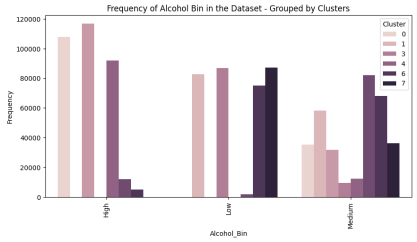
3.2.4 Feature analysis - categorical

Initially, the newly created **'alcohol_bin'** feature appears to have a uniform distribution across the dataset (fig. 7). However, a nuanced pattern emerges when considering the distribution within 'product_clusters'. Clusters 6 and 7 display a strong presence in the low alcohol category, with decreasing representation in the medium category. Conversely, the remaining clusters exhibit a predominant presence in the high alcohol category, followed by low, and then medium.

As with several other features examined, the overall distribution of the **'Color'** feature initially appears relatively uniform (fig. 8), more on the analysis

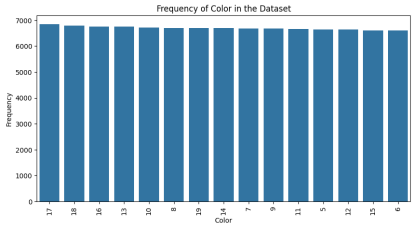


(a) Alcohol Bin

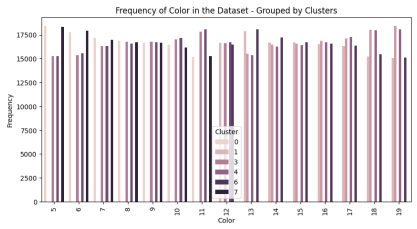


(b) Alcohol Bin Product

Figure 7: Alcohol Bin comparison



(a) Color



(b) Color Product

Figure 8: Color Comparison

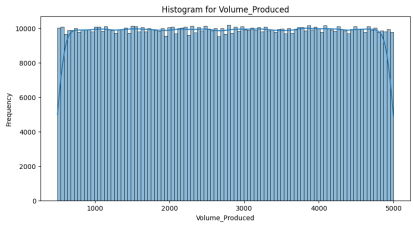
of the color on the last section (4). However, an already seen pattern emerges when we incorporate the 'product_clusters' information. Clusters 1 and 3 appear exclusively at color values above the midpoint, while clusters 7 and 0 reside solely in the lower half of the color range. Interestingly, the remaining clusters exhibit a more balanced presence across the color spectrum.

Examination of the categorical features **'Beer_Style', 'SKU', and 'Location'** revealed no discernible patterns in their overall distributions. Furthermore, incorporating cluster information did not uncover any noticeable trends or associations between these categorical features and either product or production clusters.

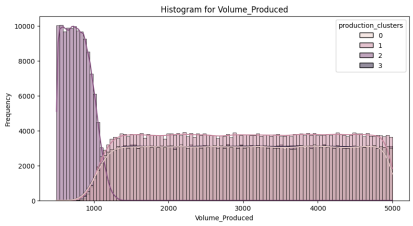
3.2.5 Feature analysis - continous

As to several features analyzed earlier, other continuous features (excluding volume produced and USD per liter) reveal recurring patterns within their frequency distributions across both production and product clusters. Regarding production clusters, clusters 0 and 3 consistently exhibit low frequencies across the analyzed features. Cluster 1 displays markedly high and consistent frequencies, while cluster 2 consistently shows the lowest frequencies.

Turning to product clusters, we observe that clusters 5, 0, and 7 consistently exhibit the highest frequencies across the features. The remaining clusters demonstrate slightly lower, but still high, frequencies.

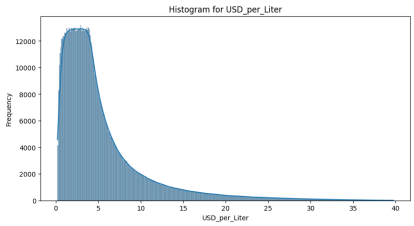


(a) Volume Produced

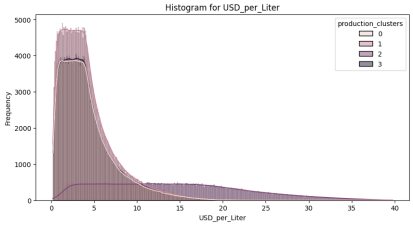


(b) Volume Produced production

Figure 9: Volume produced comparison



(a) USD per Liter



(b) USD per Liter Production

Figure 10: USD per Liter Comparison

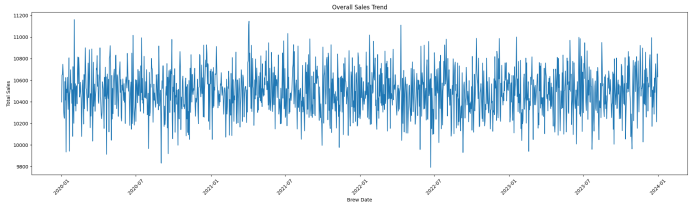
Figures 9a and 9b elucidate the distribution of production volume across clusters. Cluster 2 is characterized by a predominance of low-volume production. Conversely, cluster 1 exhibits consistent production volumes above the 1000 liter mark. Clusters 0 and 3 display similar production volume distributions between each other.

Analysis of the 'USD per liter' feature reveals distinct price distribution patterns across production clusters (fig. 10a and fig. 10b). Clusters 0 and 3, exhibiting similar price distributions, appear to do so as a consequence of their higher production volumes. Conversely, cluster 2, despite its smaller production volumes, achieves consistently higher prices across its production cycle. Cluster 1 also displays a tendency towards lower prices, likely influenced by its consistently higher production volumes

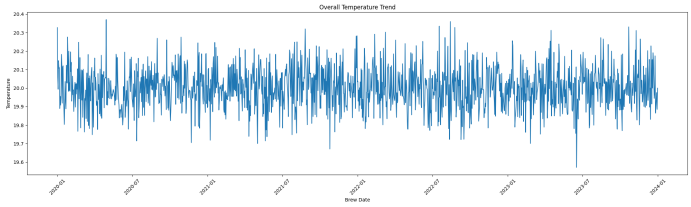
4 Conclusions and dataset evaluation

4.1 Further analysis approaches and Inconsistencies

The analysis reveals a general lack of strong correlations among the features within this dataset. This suggests the need for a more in-depth exploration in several directions. Firstly, a focused analysis on the characteristics of individual production batches could uncover nuanced patterns or relationships not evident at the broader dataset level. Secondly, the availability of production dates opens



(a) Overall Sales Trend



(b) Overall Temperature Trend

Figure 11: Trends

the possibility of conducting a time series analysis to investigate potential trends, seasonality, or temporal dependencies within the data (fig. 11a and fig. 11b)

The observed inconsistencies in the relationship between 'Color' and 'Beer_Style' warrant particular attention and underscore the importance of thorough error analysis and documentation. As illustrated in Fig.12, even beer styles with widely divergent characteristics in terms of alcohol content and fermentation time (based on industry knowledge) exhibit strikingly similar color distributions. This counterintuitive finding challenges common assumptions and highlights the need for a deeper investigation.

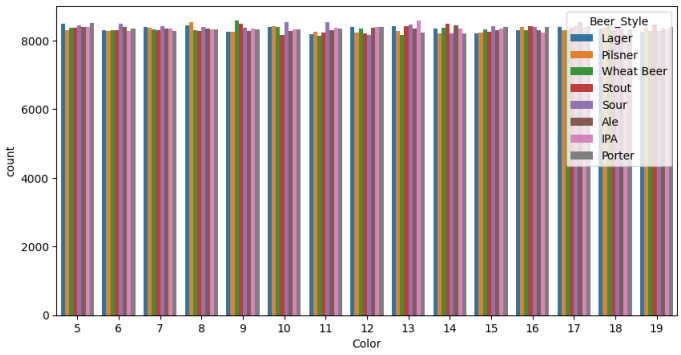
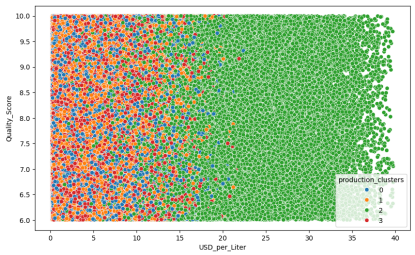
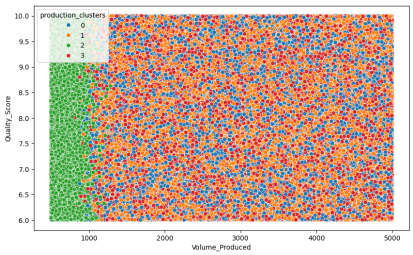


Figure 12: Beer style by color



(a) USD per liter - quality score for production



(b) Volume Produced - quality score for production

Figure 13: Quality scores distribution

4.2 Outlier Detection and distribution analysis

Various dtatistical outlier detection techniques were applied to the dataset. These included visualization using Q-Q plots and methods like the modified z-score test with adaptive thresholds. Unfortunately, these techniques yielded limited results, suggesting a lack of significant outliers within the data.

Moreover, statistical tests for skewness and kurtosis revealed that the feature distributions were generally flat, with no extraordinary patterns observed.

4.3 Quality control and Loss analysis

In an attempt to identify trends associated with high losses per batch and lower-than-usual quality scores, a comprehensive regression and classification analysis was undertaken. Several algorithms were applied to both the original dataset and the clustered data. Unfortunately, the resulting models demonstrated sub-optimal performance, yielding predictions that lacked sufficient accuracy to provide meaningful insights. For that they are not included at all inside the report.

Distribution plots were used to further examine potential relationships between 'Losses' and other loss-related features like 'Brewhouse Efficiency' and 'Brewing Efficiency'. This analysis reinforced the initial findings, demonstrating the unrelated nature of these parameters within the dataset.

Turning to the quality control task, a similar investigation was conducted with a focus on potential patterns related to the 'Quality Score' (fig 13). While this analysis revealed some minor patterns thanks to the production clustering, these results were not substantial enough to provide robust insights.