Analytics Startup Plan

Synopsis: ***This document provides a high-level walkthrough of the activities required to guide completion of the analysis.***

Project	*Brewery Recipe Optimization*
Requestor	*Max Causso*
Date of Request	*July 14,2025*
Target Quarter for Delivery	*Q3 2025*
Epic Link(s)	*Cloud-based Analytics Platform for Brewery Industry*
Business Impact	*The project aims to establish an end-to-end data analytics pipeline to uncover valuable insights from a rich dataset of beer recipes. By leveraging cloud infrastructure (Azure), predictive modeling (Python), and data visualization (Power BI), we will enable data-driven decisions for brewers, supporting product innovation, operational efficiency, and market alignment.*

1.0 Business Opportunity Brief

The craft brewing industry is characterized by high variability in recipe design, inconsistent product quality, and a lack of data-driven decision-making. Small and medium-sized breweries often rely on intuition or anecdotal experience when developing new beer recipes. With thousands of available ingredients and combinations, this trial-and-error approach results in inefficiencies, wasted resources, and missed market opportunities. The challenge is to provide brewers with reliable analytical insights that help them optimize recipe creation and predict product characteristics using historical data.

Importance for the Business & Impact:
Solving this problem is critical for enabling a more scientific and scalable approach to beer development. By leveraging a large dataset of historical recipes and modern data analytics, breweries can:

- Reduce production risk by identifying proven ingredient combinations.
- Accelerate time-to-market by narrowing experimentation to high-potential recipes.
- Align product offerings with consumer trends and preferences through data-backed insights.
- Improve overall consistency and quality in brewing operations.

This project will deliver tangible business value in the form of cost savings, process efficiency, product innovation, and enhanced customer satisfaction. It also lays the foundation for a repeatable data analytics framework that can be scaled or customized for other beverage types or regional brewing patterns.

<u>The specific ask:</u>

This project seeks to analyze a historical dataset of over 70,000 beer recipes to identify the most influential factors driving beer characteristics such as alcohol content (ABV), bitterness (IBU), and style categorization. The project will build a fully functional BI pipeline that ingests, models, and visualizes these insights using tools such as Python, Azure SQL Database, and Power BI. Final outputs will include actionable recommendations, an interactive dashboard, and a scalable architecture to support future data expansion.

1.1 Supporting Insights

- Trends: Rise in craft beer demand, consumer preference for niche styles, increasing use of data in brewing.
- Competitors: Platforms like BrewFather or BeerSmith help with recipe formulation but don’t leverage large-scale data analysis.
- Key Messages: Data-driven brewing decisions improve consistency, innovation, and reduce waste.
- Market Share: The craft brewing industry represents over 25% of beer market in North America, with strong regional variance.

1.2 Project Gains

- Revenue potential: By identifying optimal ingredient combinations and styles with higher potential, brewers can allocate resources more efficiently and reduce product development risk.
- Operational efficiency: Predictive insights will reduce trial-and-error in recipe formulation, enabling faster product launches and better alignment with market trends.
- Customer experience: Delivering beers that are more likely to meet consumer preferences enhances customer loyalty and brand differentiation.
- Strategic alignment: This project supports a broader goal of embedding analytics into product design processes and lays the foundation for future automation and recommendation systems.
- Risk of inaction: Without leveraging data, brewers may continue investing in low-performing recipes, lose market relevance, and face higher production costs due to inefficiencies.

2.0 Analytics Objective

Primary Questions:

- What are the most common ingredient combinations across beer styles?
- Can we predict alcohol by volume (ABV) or bitterness (IBU) from recipe characteristics?
- Are there natural clusters of beer recipes that could inform innovation pipelines?
- How do brewing variables (boil time, hops count, fermentable) influence product outcomes?

Hypotheses:

- Certain hop and malt combinations correlate with higher ABV and IBU.
- Recipe complexity (number of ingredients) may be linked to style classification.
- Boil time and fermentation duration are critical factors in flavor intensity.

2.1 Other related questions and Assumptions:

- Missing values in key fields (e.g., ABV or IBU) may affect modeling accuracy.
- Recipes are assumed to be valid representations, though they may vary in quality or completeness.
- Style classification is used as a proxy for consumer preference due to the lack of direct rating data.

2.2 Success measures/metrics

- Completion of a scalable ETL process using Azure Data Factory and Azure SQL
- Deployment of a fully interactive Power BI dashboard with dynamic filtering and drilldowns
- Accuracy metrics for predictive models (e.g., R^2 for ABV regression, accuracy for style classification)
- User engagement: dashboard usage by end-users and stakeholders
- Delivery of final insights and report by the defined deadline (Q3 2025)

2.3 Methodology and Approach

Type of Analysis: Linear regression, decision trees, clustering (K-Means), and exploratory data analysis (EDA). The initial approach will be to use unsupervised clustering techniques to identify natural groupings of beer recipes based on key variables such as alcohol content (ABV), bitterness (IBU), fermentation methods, and ingredient combinations. I will also apply regression analysis to predict key outcome variables (e.g., ABV) and decision trees to determine which variables (e.g., hop type, grain composition, yeast family) are most influential in style classification or in estimating bitterness. Additional statistical techniques such as correlation analysis and distribution modeling will support variable importance and normalization.

Methodology: Key questions from the ‘Analytics Objective’ section will be addressed in sequential order, as outlined in the ‘5.0 Timelines and Deliverables’ section.

I will start by performing data cleansing and exploration to address missing values and standardizing ingredient names. Once a clean base is prepared, I will create derived variables such as total number of hops per recipe, percentage of fermentable, and categorical indicators for common yeast strains.

I will then segment the dataset using K-Means clustering to reveal patterns and recipe archetypes. Following this, I will run a linear regression model to predict alcohol content (ABV) based on features like grain bill size, boil time, and hop additions. For classification tasks, I will implement a decision tree model to determine which ingredients and brewing parameters are most predictive of beer style. These models will be cross-validated and refined to ensure generalizability.

I may repeat the classification and regression analyses on specific subsets (e.g., top 5 most common beer styles) to determine whether variable importance or accuracy shifts when focusing on narrower recipe families.

Output: The output will consist of a consolidated report with key findings, including clustering profiles, predictive model summaries, and ingredient influence rankings. Strategic recommendations will be developed for brewers looking to optimize recipes, highlighting which variables to prioritize based on their impact. In addition, a fully interactive Power BI dashboard connected to Azure SQL will be delivered, allowing users to explore trends, test ingredient combinations virtually, and gain actionable insights to guide future product development.

3.0 Population, Variable Selection, considerations

- **Audience / Population Selection:** The analysis is primarily targeted at independent craft brewers, product managers, and analytics or R&D teams within beverage manufacturing companies who are seeking to innovate recipe design, optimize brewing processes, or explore data-driven decision-making.

- **Observation Window:** The dataset is static and does not contain temporal variables such as brew date, release year, or consumption period. As such, the analysis assumes time invariance and interprets recipe patterns as cross-sectional rather than time-series data.

- **Inclusions:** Only records with complete and valid values for critical fields (e.g., ABV, style, primary ingredients) are retained for analysis. Inclusion criteria ensure sufficient quality for statistical modeling and clustering.

- **Exclusions:** Recipes with excessive missing data, unparseable ingredient formats, or invalid units are excluded. Additionally, redundant records and duplicates (e.g., slight variations of the same base recipe) may be filtered based on feature similarity.

- **Data Sources:** The dataset originates from Kaggle's public "Beer Recipes" CSV file, which aggregates tens of thousands of user-submitted recipes across various beer styles, regions, and brewers. It includes structured and semi-structured fields.

- **Audience Level:** The data is modeled at the recipe level, meaning each row represents a unique beer formulation with its own combination of ingredients and brewing instructions.

- **Variable Selection:** Core fields used in the analysis include: style, abv, ibu, boil_time, hops, fermentables, yeast, grain_bill, batch_size, and color.

- **Derived Variables:** New features engineered for analysis include:

    - Hop count: Number of unique hops used per recipe

    - Fermentables ratio: % of fermentable sugars in the total grain bill

    - Ingredient diversity index: Count of distinct ingredient types

    - Style indicator dummies: One-hot encoding of major beer families

    - Bitterness-to-alcohol ratio: A derived intensity metric (IBU/ABV)

- **Assumptions and Data Limitations:**
  - The dataset lacks direct user ratings or consumption metrics, so popularity is inferred through indirect proxies such as style frequency.
  - Ingredient naming conventions are not standardized; similar ingredients may appear under different labels (e.g., "Cascade hops" vs. "Cascade"). NLP-based preprocessing and regex cleaning are applied to mitigate this issue.
  - Beer style classifications may vary between users, introducing potential noise in modeling efforts. This subjectivity is acknowledged and may be addressed by aggregating similar styles or focusing on the most consistent categories.
  - There is no indication of recipe success in production or market performance, so insights are oriented toward formulation potential rather than commercial outcome.

4.0 Dependencies and Risks

Risk	Likelihood (based on historical data)	Delay (based on historical data)	Impact
Inconsistent ingredient naming	Medium	Medium	High - Use regex/text preprocessing to group similar ingredients.
Missing values in key fields (e.g., ABV)	High	Medium	Medium - Impute where reasonable, exclude extreme cases.
No rating/popularity field	Low	Low	Medium - Use frequency of style as a proxy, explore clustering.
Limited storage access	Low	Low	Low - Azure SQL and OneDrive secured.
Generalized model may lack accuracy for specific styles	Medium	Low	Medium – A universal prediction model may underperform compared to style-specific models. Where feasible, specialized models may be developed for the most frequent beer styles (e.g., IPA, Stout) to improve accuracy.

5.0 Deliverable Timelines

Item	Major Events / Milestones	Description	Scope	Days	Date
1.	Kick-off / Formal Request	Project initiation: alignment with academic advisor and confirmation of overall objectives, scope, methodology, tools (Python, Azure, Power BI), and dataset source (Kaggle CSV).	All	-	July 15
2.	Assessment / Triage	Deep dive into the raw dataset: evaluate structure, completeness, missing values, inconsistencies (e.g., ingredient naming), and overall suitability for modeling and BI. Initial data dictionary draft.	Medium	2	July 17
3.	Prioritization	Formal identification and documentation of key research questions, hypotheses, target variables (ABV, IBU, style), and analytical success metrics (e.g., model accuracy, dashboard usability).	Medium	1	July 18
4.	Data Exploration & Analysis • Issues with duplicates • Missing/Invalid values	Execution of full EDA: data cleaning, treatment of missing/null values, removal of duplicates, and generation of derived fields (ingredient counts, bitterness-to-alcohol ratio, fermentation ratios). Start of model-ready dataset creation. • Detect and resolve near-duplicate recipes using similarity rules (e.g., same name, ingredients, style). • Apply thresholds to remove low-quality records and/or impute reasonable defaults where applicable.	High	3	July 21

5.	Story Board 1	Draft version of the Power BI dashboard: first version of visuals with ABV, IBU, style distributions, ingredient usage, and clustering previews. Layout structure and navigation tested.	Medium	2	July 23
6.	QA Output	Internal quality review of data integrity, feature correctness, model interpretability, and dashboard filters/UX. Ensure consistency between backend logic and front-end visuals.	Medium	2	July 25
7.	Internal Team Presentation	Present preliminary findings, model performance, and dashboard navigation to internal reviewers or academic supervisor. Collect feedback for refinement.	All	1	July 26
7.	Go / No-Go	Review critical issues raised. Based on feedback, decide whether to proceed to final delivery or iterate further on models and visualizations. Formal checkpoint.	All	1	July 27
9.	Story Board 2	Develop the final dashboard incorporating all feedback. Add any remaining visualizations (e.g., style prediction trees, cluster profiles, variable importances) and finalize KPI cards.	Medium	3	July 30
10.	Pilot	Controlled release: simulate stakeholder interaction with dashboard. Validate interactivity, responsiveness, and interpretation of outputs. Measure engagement and identify usability issues.	Medium	2	August 1
11.	Delivery & Sign-off	Final handover: submit cleaned dataset, models (if applicable), documented methodology, key insights, dashboard file and link (Power BI), and executive summary. Ensure all files are versioned and archived.	All	9	August 10