



# **GA DSI 26**


## **Project 3:**

# **Wine and Beer**

By: Lim Zhi Yong

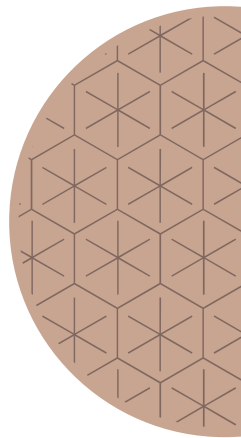


# Task:

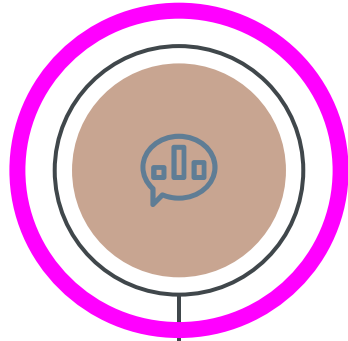
- Understand consumer patterns
  - Identify if the consumer wants winemaking or homebrewing info
  - Train model with subreddit posts
- 

# Data Description

- 1,000 posts from each subreddit
  - r/winemaking
  - r/homebrewing
- Cleaned punctuation, stopwords, delimiters
- Considered both unigrams and bigrams



# Model Workflow



## Data Cleaning

- Missing values
- Creating new features
- Choosing features



## Modelling

- Building models
- Scoring models



## Testing

- Kaggle testing on test dataset



## Recommendations

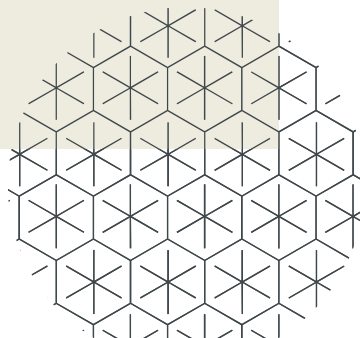
- Important features

# Missing Values

There are different types of missing values:

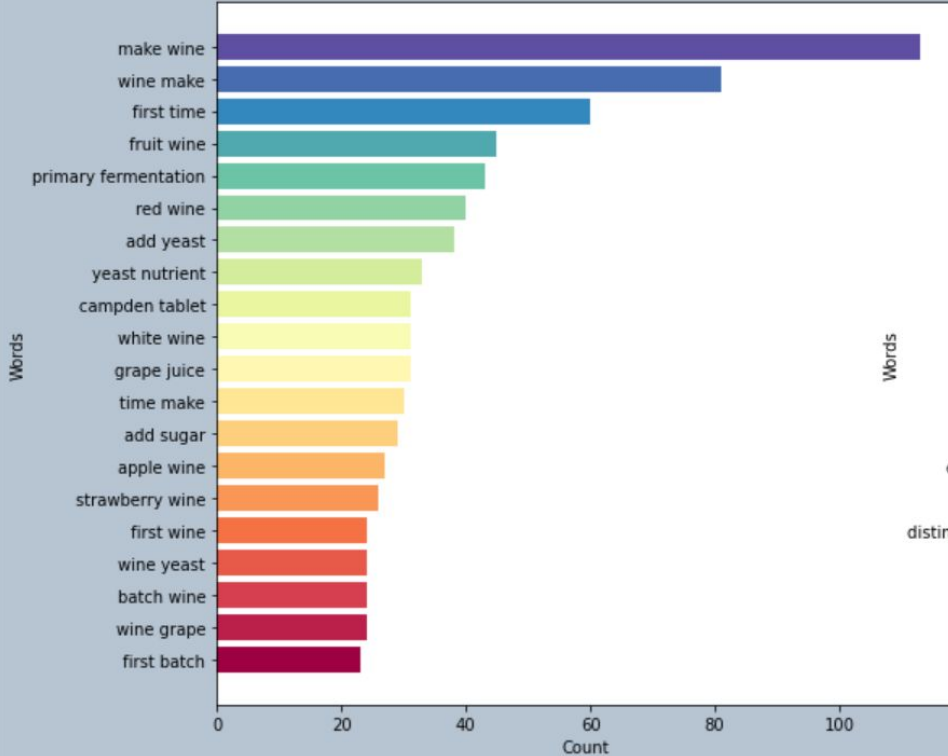
- Duplicate posts were removed
- Null and removed texts were replaced with the empty string
- One deleted post was miscategorized, we removed it

1969 rows left

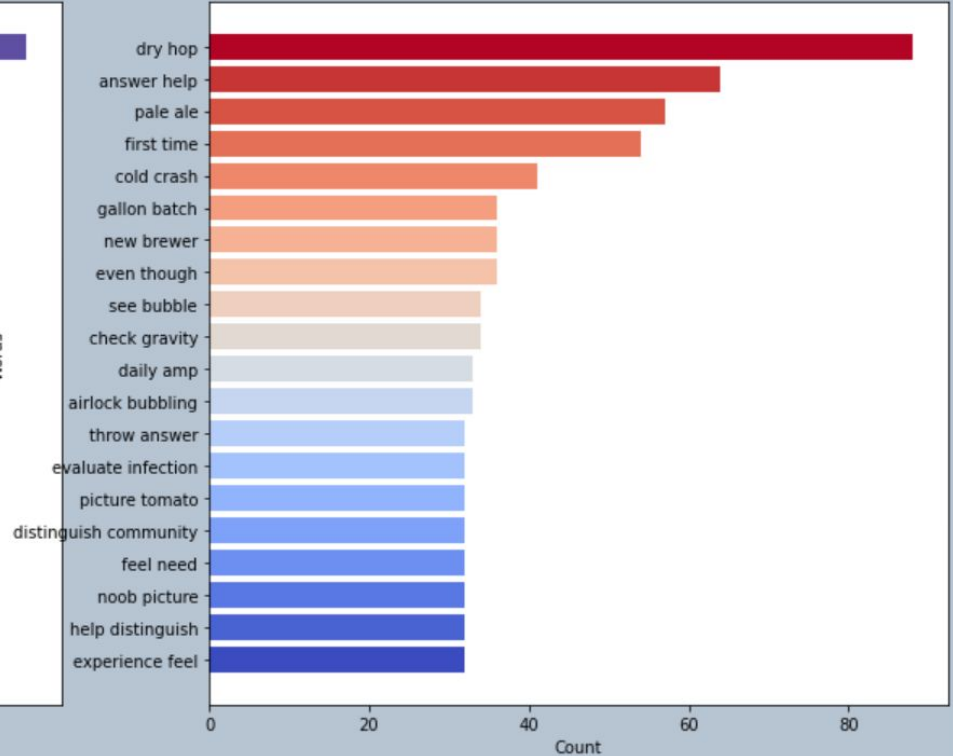


## Plots of most frequent bigrams in the subreddits

Top 20 words in winemaking subreddit titles

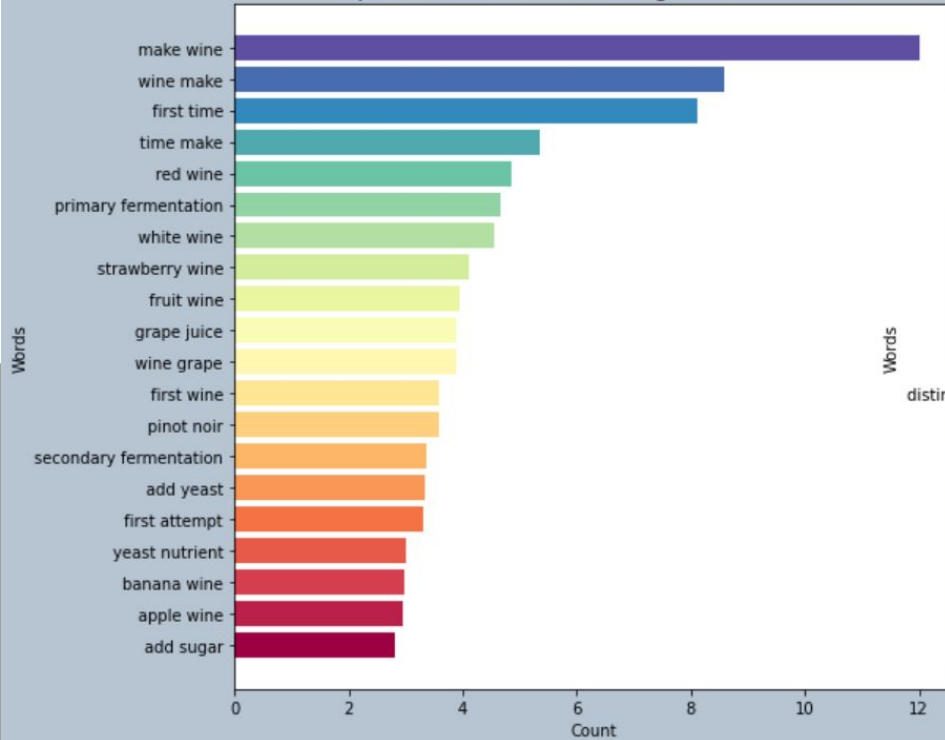


Top 20 words in homebrewing subreddit text

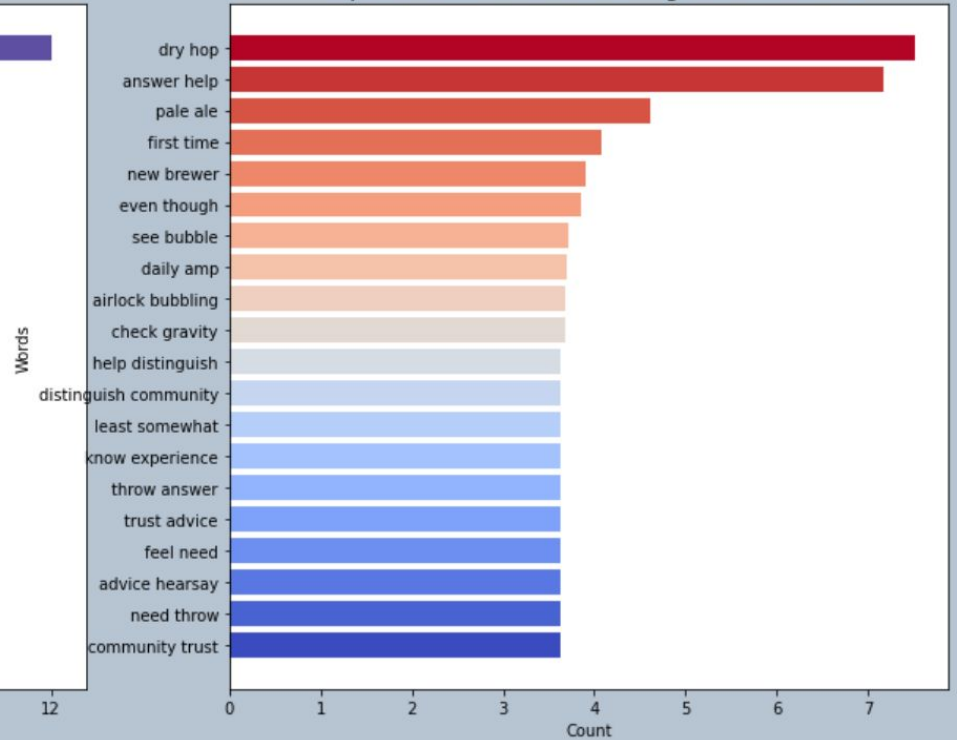


## Plots of most frequent tf-idf bigrams in the subreddits

Top 20 words in winemaking subreddit titles



Top 20 words in homebrewing subreddit text

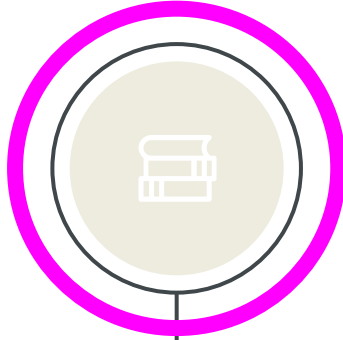


# Model Workflow



## Data Cleaning

- Missing values
- Creating new features
- Choosing features



## Modelling

- Building models
- Scoring models



## Testing

- Kaggle testing on test dataset



## Recommendations

- Important features



# Modelling

- ❖ 8 models:
  - Logistic regression
    - Count
    - Tf-idf
  - KNN classifier
    - Count
    - Tf-idf
  - Naïve bayes
    - Count
    - Tf-idf
  - Random forest
    - Count
    - Tf-idf

## Scoring

- ❖ ROC-AUC to determine best models
- ❖ F1 score to compare baseline score
- ❖ Accuracy to determine whether overfit

# Model Workflow



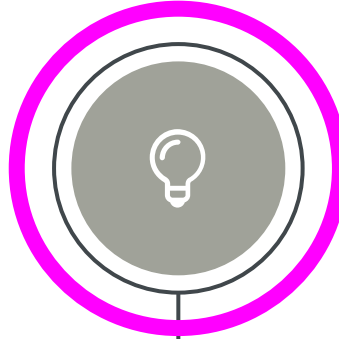
## Data Cleaning

- Missing values
- Creating new features
- Choosing features



## Modelling

- Building models
- Scoring models



## Testing

- Kaggle testing on test dataset

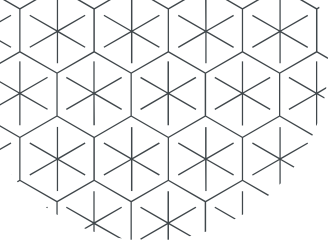


## Recommendations

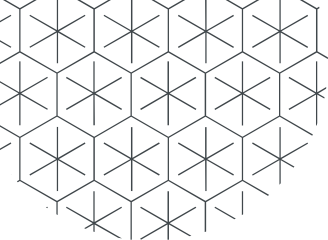
- Important features

The background features several abstract, organic shapes in muted colors: a large brownish-tan shape on the left, a large light beige shape in the center, and a greyish-green shape on the right. A small, stylized black line drawing of a leafy branch is positioned at the top center. The text is centered over the beige shape.

**Logistic Regression  
performed the best**



**0.97 ROC-AUC**  
**0.92 accuracy**



models	vectorizer	accuracy score	auc score
Logistic Regression	count	0.897	0.961
Logistic Regression	tf-idf	0.917	0.975
KNN Classifier	count	0.720	0.851
KNN Classifier	tf-idf	0.580	0.655
Naïve Bayes	count	0.789	0.925
Naïve Bayes	tf-idf	0.789	0.925
Random Forest	count	0.890	0.964
Random Forest	tf-idf	0.888	0.966

# Model Workflow



## Data Cleaning

- Missing values
- Creating new features
- Choosing features



## Modelling

- Building models
- Scoring models



## Testing

- Kaggle testing on test dataset



## Recommendations

- Important features

# Best Model

## Features

- Wordnet lemmatizer
- Tf-idf vectorizer
- Logistic regression with ridge penalty

## Limitations

- Spell check before lemmatizing
- Slightly overfit, could remove more stopwords





# Thanks

CREDITS: This presentation template was created  
by **Slidesgo**, including icons by **Flaticon**,  
infographics & images by **Freepik**

Please, keep this slide for the attribution