A

Web Technology Mini Project Report submitted to Savitribai Phule Pune University, Pune

“Food Restaurant Data Analysis”

A

In partial Fulfillment for the awards of Degree of Engineering in Computer Engineering

Submitted by

Mr. Ankit Dadmal T1902404248

Ms. Aditi Jadhav T1902404302

Under the Guidance of

Prof. Rubi Mandal

Department of Computer Engineering

2024-25

DYPDPU
Dr. D. Y. Patil Unitech Society's
Dr. D. Y. Patil Institute of Technology

Certificate

This is to certify that,

Mr. Ankit Dadmal T1902404248
Ms. Aditi Jadhav T 1902404302

have successfully completed the Mini project entitled "Food Restaurant Data Analysis "under my guidance in partial fulfillment of the requirements for the Third Year of Engineering in Computer Engineering under the Savitribai Phule Pune University during the academic year 2024-2025

Date: …….

Place: …….

Prof. Rubi Mandal

Project Guide

Dr. Vinod Kimbahune

HOD

Dr. Nitin Sherje
Principal

Acknowledgement

With deep sense of gratitude, we would like to thank all the people who have lit our path with their kind guidance. We are very grateful to these intellectuals who did their best to help during our project work.

It is our proud privilege to express a deep sense of gratitude to **Prof. Dr. Nitin Sherje** Principal of for his comments and kind permission to complete this project. We remain indebted to **Dr. Vinod Kimbahune,** H.O.D. Computer Engineering Department for his timely suggestion and valuable guidance.

The special gratitude goes to **Prof. Rubi Mandal** excellent and precious guidance in completion of this work. We thanks to all the colleagues for their appreciable help for our working project. With various industry owners or lab technicians to help, it has been our endeavor throughout our work to cover the entire project work.

We are also thankful to our parents who provided their wishful support for our project completion successfully. And lastly, we thank our all friends and the people who are directly or indirectly related to our project work.

Mr. Ankit Dadmal
Ms. Aditi Jadhav

1. Abstract

The exponential growth of the food industry and the increasing reliance on digital platforms for discovering and ordering food have created a need for insightful data analysis to better understand food preferences, regional culinary diversity, and cooking patterns. This mini project focuses on the comprehensive analysis of a food restaurant dataset sourced from Kaggle, which includes features such as dish name, ingredients, diet type, preparation time, cooking time, flavour profile, course, state, region, and image URL.

The objective of the project is to apply data science and business analytics techniques to uncover patterns, correlations, and trends within the dataset. A variety of visualizations have been created, including word clouds, line plots, bar charts, pie charts, scatter plots, box plots, violin plots, and heat maps. These visual representations provide valuable insights into ingredient popularity, regional dish distribution, preparation and cooking time variations, and flavour preferences across India.

By employing Python and its data analysis libraries (Pandas, Matplotlib, Seaborn, and Word Cloud), the project demonstrates how exploratory data analysis (EDA) can be used to generate actionable insights in the food domain. The findings can support restaurant owners, food aggregators, and culinary researchers in optimizing their offerings, improving customer satisfaction, and making data-driven decisions.

Table of Contents

List of Abbreviations

Sr. No.	Abbreviation	Full Form
1	DSBDA	Data Science and Business Data Analytics
2	EDA	Exploratory Data Analysis
3	CSV	Comma-Separated Values
4	UI	User Interface
5	ML	Machine Learning
6	Fig	Figure
7	RAM	Random Access Memory
8	SSD	Solid State Drive
9	IDE	Integrated Development Environment
10	HTML	HyperText Markup Language
11	CSS	Cascading Style Sheets
12	JS	JavaScript
13	URL	Uniform Resource Locator
14	KPI	Key Performance Indicator

List of Figures

2. Introduction

The food industry is one of the largest and most rapidly expanding sectors globally, reflecting the growing demand for diverse culinary experiences, evolving consumer preferences, and increased reliance on digital platforms for food ordering and delivery. In an age where competition is fierce, restaurants and food businesses must adopt innovative approaches to stay competitive, meet consumer expectations, and optimize operational efficiency. A key to success in this environment is the integration of data-driven decision-making through data analytics.

The ability to analyze large datasets of food-related information, such as ingredients, preparation times, cooking times, and regional dish variations, offers significant opportunities for gaining valuable insights into food trends, customer behaviors, and menu optimization. With the increasing adoption of digital technologies, the food industry is shifting towards data-driven strategies to enhance customer satisfaction, streamline operations, and improve overall business performance.

This project aims to apply data science techniques to analyze a comprehensive food restaurant dataset, sourced from Kaggle. The dataset includes detailed information about various dishes, including their names, ingredients, dietary information (e.g., vegetarian, non-vegetarian, vegan), preparation and cooking times, flavor profiles, courses (e.g., appetizers, main dishes, desserts), and regional data. By applying various data analysis and visualization techniques, the project provides insights into the following areas:

1. **Ingredient Analysis**: Identifying the most commonly used ingredients across dishes, which can help in understanding food trends, consumer preferences, and popular combinations.

2. **Regional Dish Distribution**: Analyzing how different dishes are distributed across regions and states, uncovering regional food preferences and cultural influences in culinary offerings.

3. **Preparation and Cooking Time Analysis**: Evaluating the efficiency of dish preparation and cooking processes by analyzing the time spent on different dishes, offering insights into operational improvements and customer expectations regarding delivery times.

4. **Flavor Profile Analysis**: Investigating the flavor profiles of dishes (e.g., spicy, sweet, Savory) and how they vary across different courses and regions, providing a better understanding of consumer taste preferences.

5. **Data Visualization**: The project leverages various data visualization techniques, including word clouds, line plots, bar charts, pie charts, heatmaps, and scatter plots, to present findings in a clear and accessible manner. These visualizations help communicate complex data insights in a way that is easily interpretable for restaurant owners, chefs, and other stakeholders.

The primary objective of this project is to offer actionable insights into the food industry through data analytics, thereby enabling restaurant managers and food businesses to make informed decisions about their menus, optimize operations, improve customer experiences, and adapt to evolving trends in the culinary market.

By using a combination of Python, Pandas, Matplotlib, Seaborn, and other data analysis libraries, this project aims to explore and extract meaningful insights from the dataset. The findings can be used to enhance the efficiency and competitiveness of restaurants and food businesses, offering a data-driven approach to understanding customer demands, food trends, and regional preferences.

3. Objective of the Project

The primary objective of this project is to conduct a comprehensive analysis of a food restaurant dataset, which contains various features such as dish names, ingredients, preparation and cooking times, diet types, course types, flavor profiles, and regional data. By applying data analysis and visualization techniques, the project seeks to uncover valuable patterns, trends, and relationships that can provide actionable insights into the food industry. The analysis aims to inform restaurant owners, food businesses, and chefs about key aspects of their operations and customer preferences.

The main objectives of this project are as follows:

1. **Explore the Relationships Between Different Features**: One of the key goals of this project is to investigate how various features of the dataset are interrelated. For example, the project will examine how **ingredients** affect **preparation and cooking times**, or how the **region** impacts the popularity of certain **dishes**. By understanding these relationships, the project will provide a deeper insight into the dynamics of food preparation, customer preferences, and regional culinary trends.
    - **Example Investigation**: Are dishes with a higher number of ingredients generally associated with longer cooking times? Is there a correlation between certain regions and specific ingredients or flavor profiles?

2. **Provide Insights into the Most Common Ingredients Used in Dishes**: Another primary objective is to analyze the ingredients used across the dataset to determine the most common and widely used ingredients in food dishes. This insight can help restaurant owners and food businesses identify popular ingredients and adjust their menus to reflect consumer preferences.
    - **Example Outcome**: Which ingredients appear most frequently in dishes across all regions and courses? Are there any noticeable food trends or preferences, such as the growing popularity of plant-based ingredients or certain spices?

3. **Examine the Variation in Preparation and Cooking Times**: A critical part of the analysis is to evaluate how **preparation times** and **cooking times** vary for different types of dishes, regions, and courses. This examination will provide valuable insights into the efficiency of different dishes, the time required for their preparation, and any

potential bottlenecks in the kitchen.

- **Example Investigation**: Do **main courses** generally take longer to prepare and cook than **appetizers** or **desserts**? How do **regional variations** affect cooking and preparation times? Are there certain **dishes** that consistently have longer or shorter times across the dataset?

4. **Analyze the Distribution of Dishes Based on Flavor Profile, Course, and Regional Preferences**: This project also seeks to explore the distribution of dishes based on their **flavor profile**, **course**, and **region**. Flavor profiles, such as sweet, savoury, spicy, and sour, play a major role in the culinary experience and can reveal regional or cultural preferences. By understanding the distribution of these flavor profiles, the project will help provide recommendations for menu planning and regional customization.
    - **Example Investigation**: How do **flavor profiles** (e.g., spicy, sweet, savoury) vary across **courses** (e.g., appetizers, main courses, desserts)? Are certain **regions** more inclined toward specific flavor profiles? What courses have a greater tendency for specific flavor types (e.g., are desserts predominantly sweet, or are regional dishes more savoury)?

Additional Project Goals:

- **Visualization of Key Findings**: Throughout the analysis, various types of visualizations, such as word clouds, bar charts, line plots, scatter plots, pie charts, and heatmaps, will be employed to present the findings in an accessible manner. These visualizations will help simplify complex data and make the insights easy to interpret for decision-makers.
- **Provide Actionable Insights for Menu Optimization**: By understanding the trends, preferences, and operational efficiencies from the data, the project aims to provide actionable recommendations for food businesses. These insights can help businesses make informed decisions about menu design, ingredient sourcing, pricing strategies, and regional adaptations.

4. About the Dataset

4.1 Source of Dataset

The dataset used in this project is sourced from Kaggle. It includes information on various dishes from restaurants across India, containing attributes such as dish name, ingredients, diet type, preparation time, cooking time, flavor profile, course, state, region, and image URL. This dataset provides a comprehensive look at the food landscape in India.

4.2 Features Description

The dataset consists of the following key features:

- **Dish Name**: The name of the dish.
- **Ingredients**: A list of ingredients used in the dish.
- **Diet Type**: Whether the dish is vegetarian, non-vegetarian, or vegan.
- **Prep Time**: The time required to prepare the dish.
- **Cook Time**: The time taken to cook the dish.
- **Flavor Profile**: The type of flavor (e.g., spicy, sweet, sour) associated with the dish.
- **Course**: The course of the meal (e.g., appetizer, main course, dessert).
- **State**: The state in India where the dish originates.
- **Region**: The specific region within the state.
- **Img_URL**: A URL link to an image of the dish.

5. Tools and Technologies Used

In this project, several tools and technologies were used to analyze and visualize the restaurant dataset. Below is a detailed description of the key tools and technologies utilized:

Python

- **Description**: Python is the main programming language used in this project for data analysis and visualization. Python is widely used in the data science community due to its simplicity, readability, and extensive libraries for data manipulation and visualization.
- **Usage**: Python was employed to load, clean, and analyze the dataset, perform various data wrangling tasks, and create visualizations.

Pandas

- **Description**: Pandas is a powerful and flexible data manipulation library in Python. It is designed for working with structured data, such as CSV files and SQL databases. Pandas provide efficient tools for data cleaning, transformation, and analysis.
- **Usage**: In this project, Pandas was used for loading and manipulating the dataset, handling missing values, and preparing the data for visualization. It was essential in handling the tabular data and performing operations like grouping, filtering, and aggregating.

Matplotlib

- **Description**: Matplotlib is a widely used Python library for creating static, animated, and interactive visualizations. It provides a variety of tools for generating charts, plots, and graphs.
- **Usage**: Matplotlib was used in this project to create various visualizations, including line plots, bar charts, histograms, and scatter plots. It helped present the data analysis in a clear and visually appealing manner.

Seaborn

- **Description**: Seaborn is a Python data visualization library built on top of Matplotlib that provides a high-level interface for drawing attractive and informative statistical graphics. It simplifies the process of creating complex plots like box plots, violin plots, and heatmaps.
- **Usage**: Seaborn was used for statistical visualizations, including box plots, violin plots, and heatmaps. Its integration with Pandas made it easy to work with data frames and create aesthetically pleasing charts with minimal code.

Word Cloud

- **Description**: Word Cloud is a Python library used to generate word clouds, which visually represent the most frequent words in a dataset. Word clouds help to highlight the most common ingredients or keywords by varying the size of the words based on their frequency.
- **Usage**: Word Cloud was employed in this project to visualize the most common ingredients across the dataset. The larger the word, the more frequently it appears in the dataset, which allows for easy identification of popular ingredients.

Jupyter Notebook

- **Description**: Jupyter Notebook is an open-source web application that allows users to create and share documents containing live code, equations, visualizations, and narrative text. It is widely used in data science for exploratory data analysis and reporting.
- **Usage**: Jupyter Notebook was used to write and execute the Python code for data analysis and visualization. It provided an interactive environment for experimenting with different data manipulation techniques and creating dynamic visualizations.

VS Code

- **Description**: Visual Studio Code (VS Code) is a popular, lightweight code editor that supports a wide range of programming languages and tools. It is widely used by developers for writing, debugging, and executing code.
- **Usage**: VS Code was used for writing Python scripts and managing the overall project. It provided an integrated development environment for coding, debugging, and managing libraries, making the development process more efficient.

6. Data Preprocessing

Data preprocessing is a critical step in the data analysis pipeline as it ensures that the dataset is clean, consistent, and ready for further analysis. In this project, the data underwent several preprocessing steps to address issues like missing values, duplicates, and inconsistent data formats. Below is a detailed description of the data preprocessing steps performed.

6.1 Data Cleaning

Before performing any analysis, the dataset was thoroughly cleaned to ensure its accuracy and integrity.

- **Removing Duplicates**:
  - Duplicate rows can lead to biased or incorrect analysis. In the initial stage, the dataset was checked for any duplicate rows, and they were removed. This was done using the drop_duplicates () method in Pandas to ensure that each row in the dataset represented a unique dish.
- **Data Formatting**:
  - Inconsistent naming conventions were corrected to standardize the dataset. For instance, the ingredients, diet types, and regions were checked for any inconsistencies such as varying spellings or cases (e.g., "vegetarian" vs. "Vegetarian").
  - All text fields were converted to a consistent format using string manipulation methods like str.lower() and str.strip() to remove leading or trailing spaces and ensure uniformity across categorical variables like diet types and ingredients.

6.2 Handling Missing Values

Missing values are a common issue in real-world datasets and need to be appropriately handled to ensure accurate analysis. In this project, missing values were dealt with using suitable imputation techniques.

- **Categorical Columns**:
  - For categorical variables like diet type and region, missing values were imputed with the most frequent value (mode). This technique was chosen as it ensured that the imputation was consistent with the existing distribution of values in the dataset.
  - The fillna() method in Pandas was used to fill missing categorical values with the most frequent category.
- **Numerical Columns**:
  - For numerical columns like prep time and cook time, missing values were handled by imputing with either the mean or median, depending on the distribution of the data.
  - If the data was normally distributed, the mean was used, while for skewed data, the median was chosen to avoid outliers influencing the imputed value.
  - This ensured that the dataset remained complete while minimizing the impact of missing data on the analysis.

6.3 Data Transformation

Data transformation helps in creating new features that can provide more meaningful insights during the analysis phase.

- **Creating a Total Time Feature**:
  - To facilitate the analysis of cooking and preparation times, a new feature was created by combining prep_time and cook_time. This new feature, total_time, represented the total time required to prepare and cook each dish.
  - The transformation was done by simply adding the values of prep_time and cook_time for each dish. This new variable helped in understanding the overall time commitment for preparing each dish and was useful in visualizations such as the line plot and box plot related to preparation and cooking times.
- **Categorical Encoding**:
  - Some machine learning algorithms and data visualization tools require categorical data to be represented numerically. Thus, categorical variables like diet type, flavor profile, and course were encoded using one-hot encoding. This transformed each category into a binary column representing the presence or absence of each category.
- **Normalization of Numerical Features**:
  - In cases where numerical features had different ranges (e.g., prep time and cook time), normalization was performed to scale the values within a consistent range. This was particularly important for algorithms and visualizations that might be sensitive to varying scales.
  - Min-Max normalization was used to scale the numerical values between 0 and 1, making them comparable and ensuring that features like prep_time and cook_time could be analyzed without disproportionate influence from one feature over the other

7. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is crucial in understanding the underlying patterns and trends in the dataset. Various visualizations were created to explore the data:

7.1 Overview of Visualizations

- Word Cloud for ingredient frequency
- Line plots for prep and cook times
- Bar charts for regional dish distribution
- Pie charts for flavor profile distribution
- Scatter plots to examine relationships between cooking and prep times

7.2 Interpretation of Results

The visualizations revealed key insights such as:

- The most common ingredients across dishes.
- Regional variations in dish preferences and cooking times.
- The relationship between prep time and cook time across different dishes.

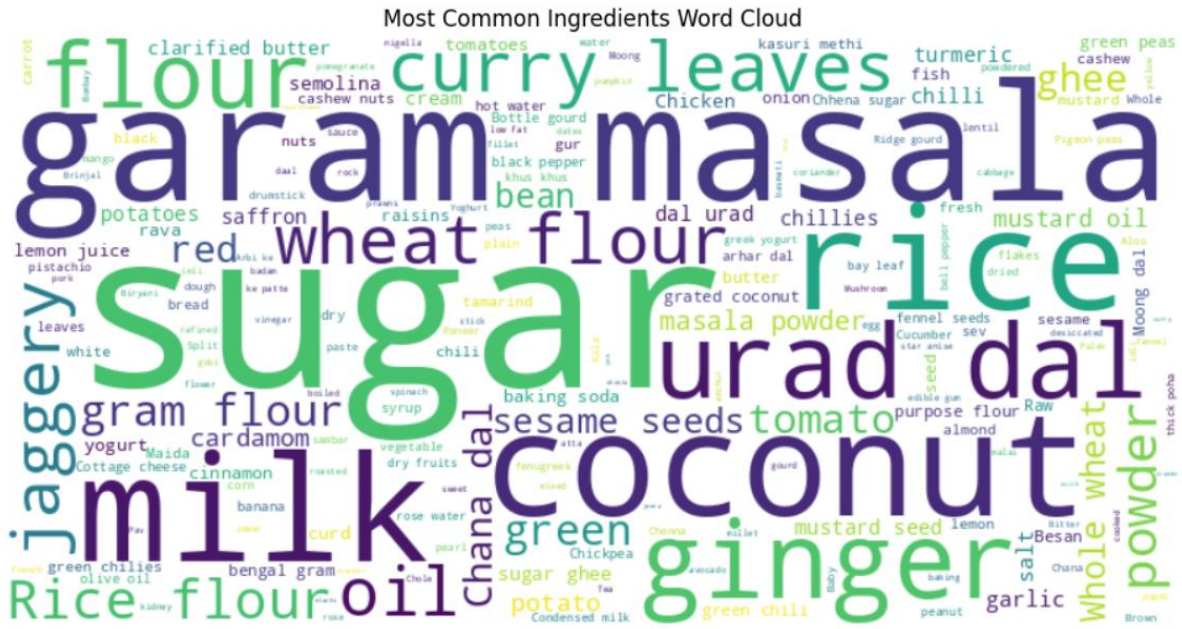
8. Visual Analysis of Food Restaurant Data

To derive meaningful insights from the dataset, several types of visualizations were created using Python libraries such as Matplotlib, Seaborn, and Word Cloud. These visualizations help to explore, analyze, and interpret the data in an intuitive and interactive way.

---

1. Word Cloud – Most Common Ingredients

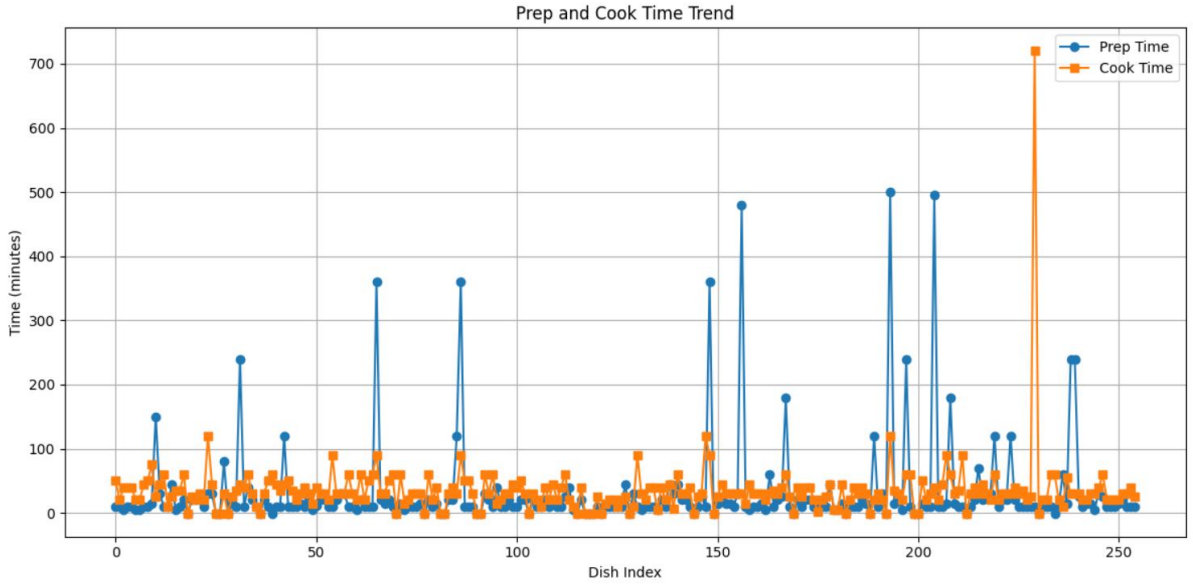
Purpose:
The word cloud represents the most frequently used ingredients in the dataset.
Insight:
Larger words indicate ingredients that appear more frequently across dishes. This helps identify common food components like "salt", "onion", "oil", etc.
Importance:
It gives a quick and visual overview of core ingredients used in Indian cuisine.



---

2. Line Plot – Prep & Cook Time by Dish Index

Purpose:
This line plot compares the preparation and cooking time across dishes indexed in the dataset.
Insight:
It shows variations in time taken for different dishes and identifies dishes with significantly high or low prep/cook times.
Importance:
Helps in understanding dish complexity and aids in kitchen planning.



3. Bar Chart – Dishes by Region

Purpose:

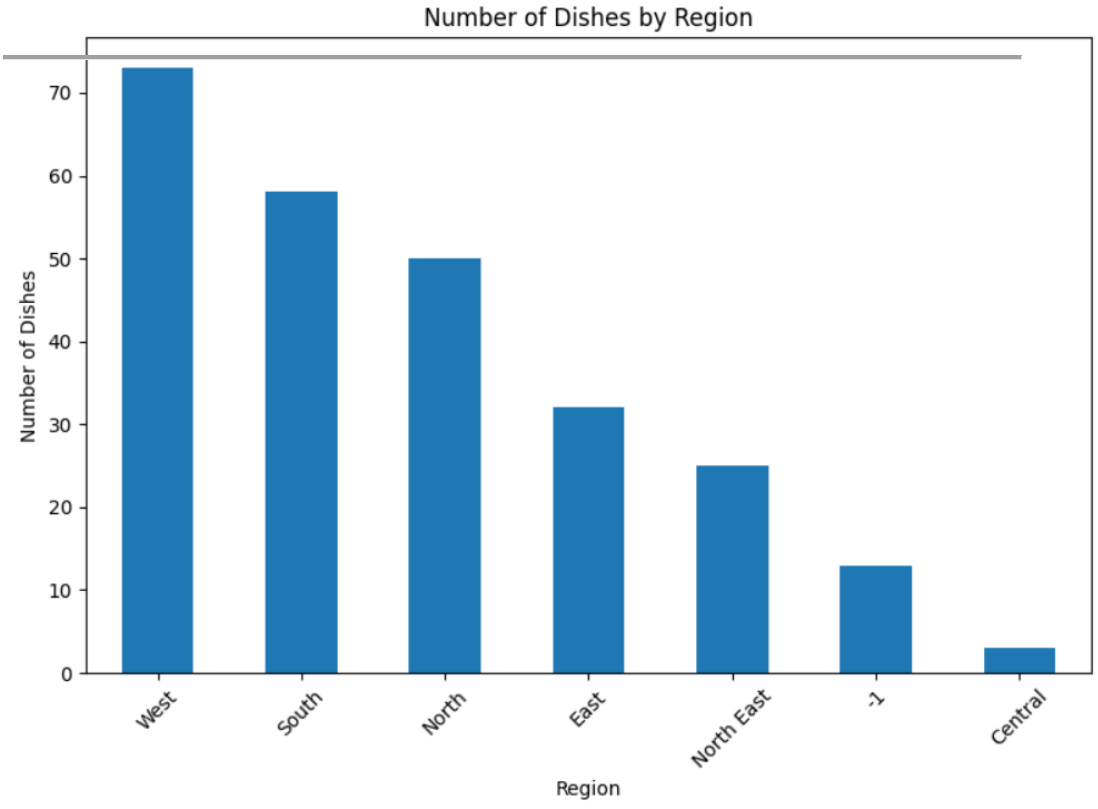
To display the number of dishes associated with each Indian region.

Insight:

It shows which regions contribute more dishes in the dataset, e.g., Northern, Southern, Western regions.

Importance:

Reveals regional diversity and popular culinary zones.



4. Horizontal Bar Chart – Dishes by Course

Purpose:

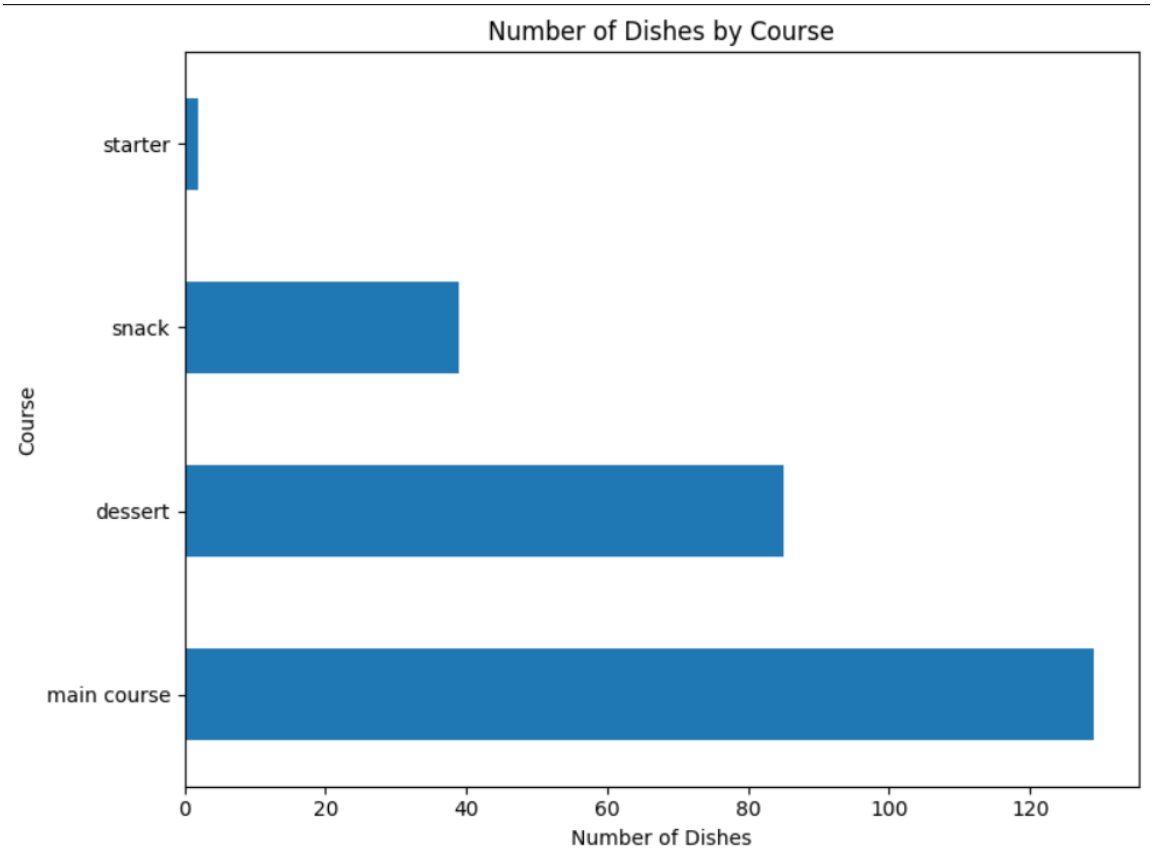
This chart shows how many dishes fall under different courses like snacks, main course, desserts, etc.

Insight:

Main course and snacks tend to dominate, while beverages and desserts are fewer.

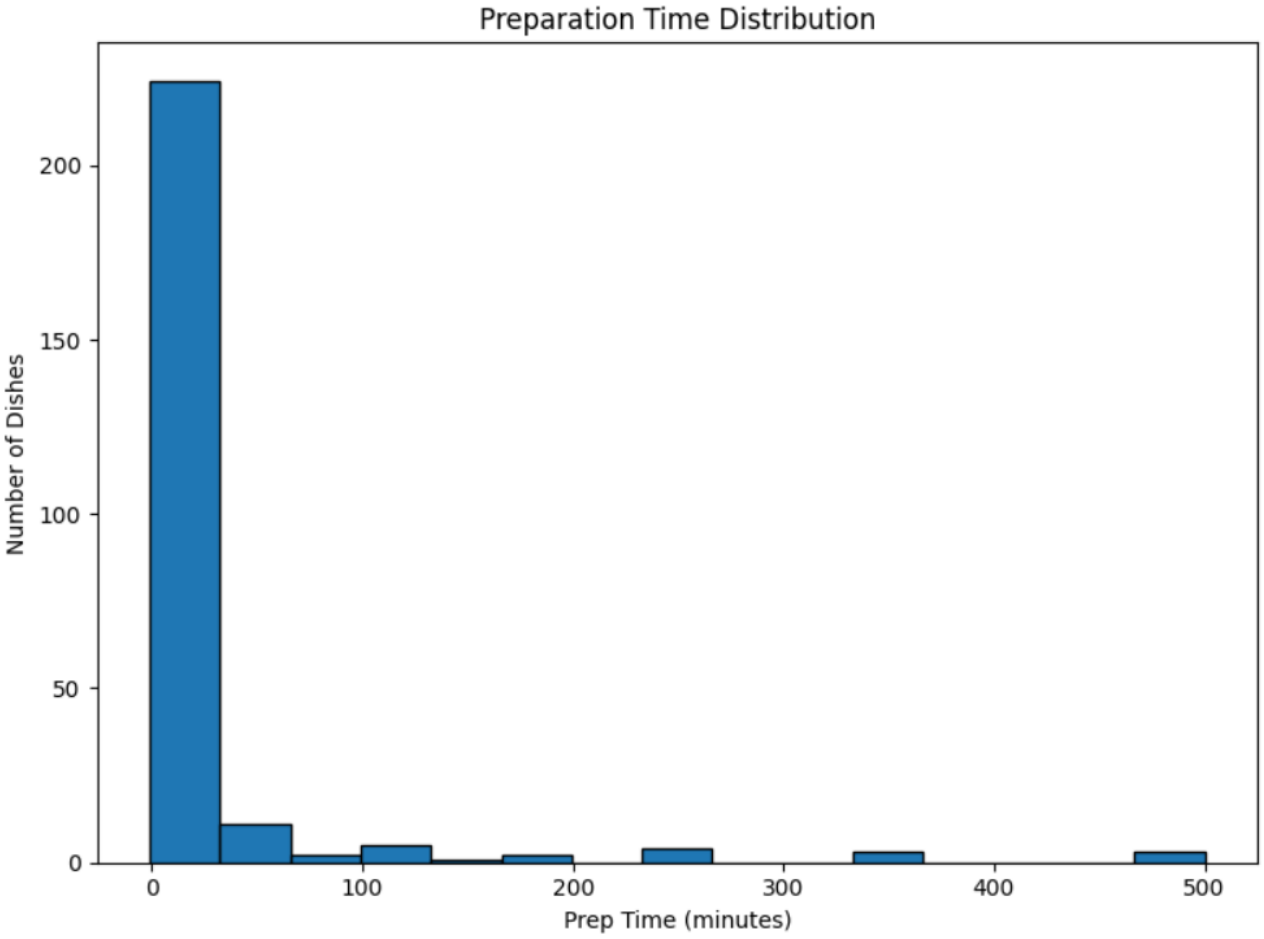
Importance:

Helps understand the dataset's structure and consumer focus.



5. Histogram – Prep Time Distribution

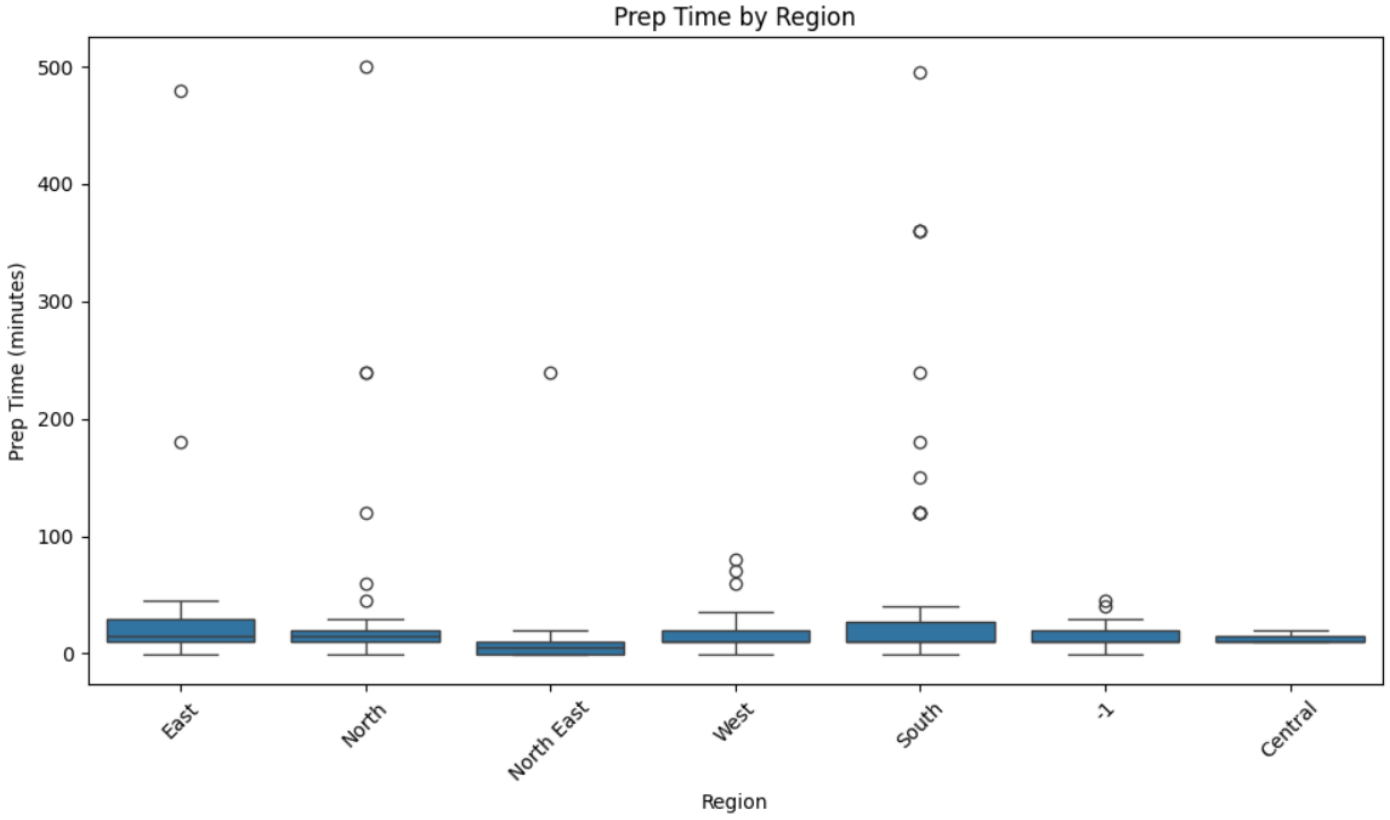
Purpose:
Shows the distribution of preparation times across all dishes.
Insight:
Most dishes have prep times between 10–30 minutes, with few taking longer.
Importance:
Highlights ease or complexity of recipe preparation.



---

6. Box Plot – Prep Time by Region

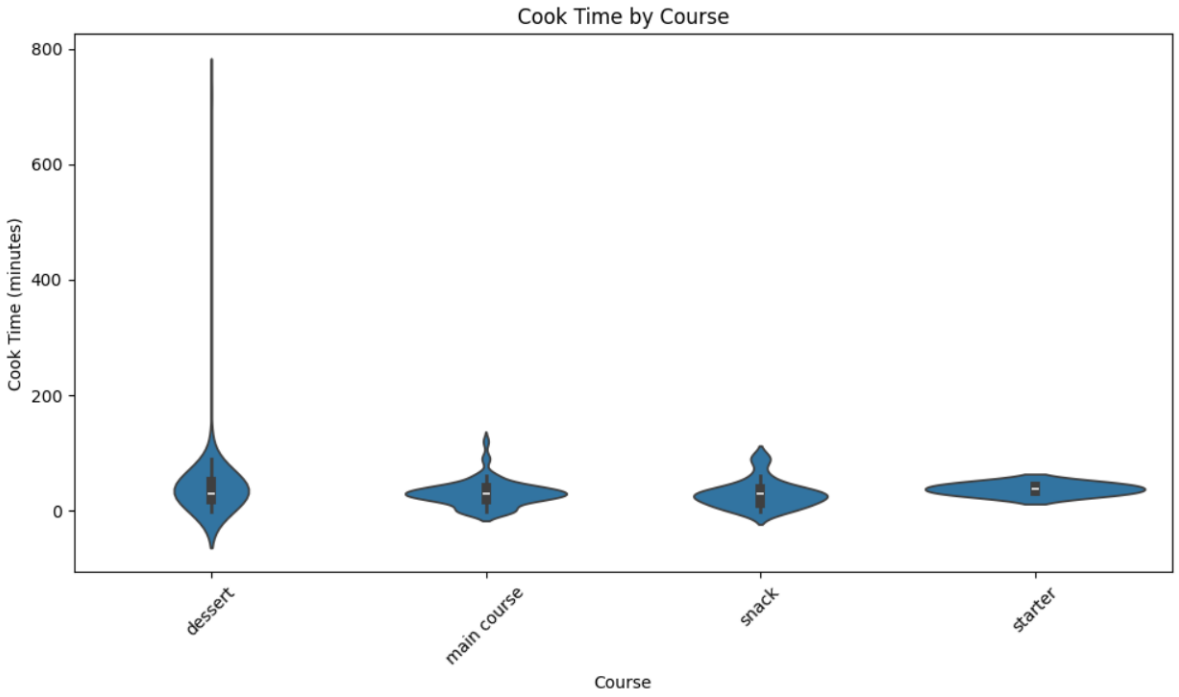
Purpose:
Compares the spread of preparation times across different regions.
Insight:
Outliers and median values help analyze which regions have more time-intensive recipes.
Importance:
Useful for analyzing regional differences in cooking complexity.



---

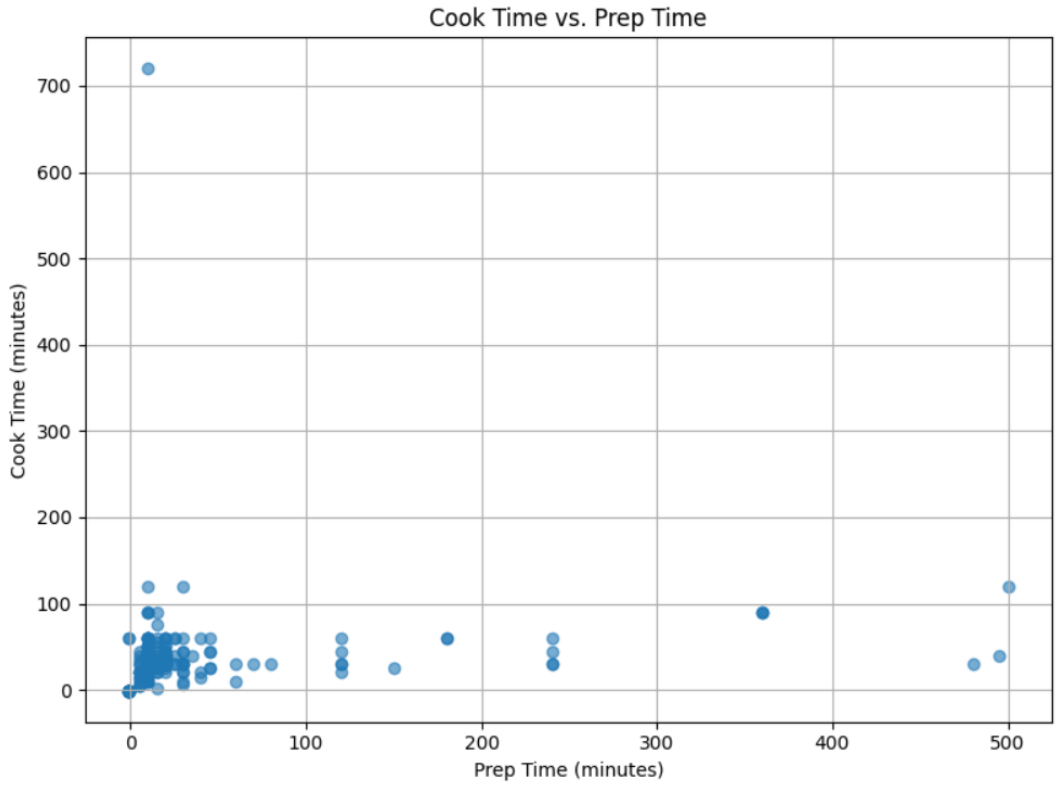
7. Violin Plot – Cook Time by Course

Purpose:
Illustrates the distribution and density of cook times by course.
Insight:
Shows which types of courses generally take more/less cooking time.
Importance:
Combines box plot and KDE plot for better time comparison.



8. Scatter Plot – Cook Time vs Prep Time

Purpose:
Plots cooking time against preparation time to examine correlation.
Insight:
Identifies whether longer prep times relate to longer cook times. Some dishes take long to prep but less to cook and vice versa.
Importance:
Useful for optimizing recipe processes.



---

9. Pie Chart – Flavor Profile Distribution

Purpose:

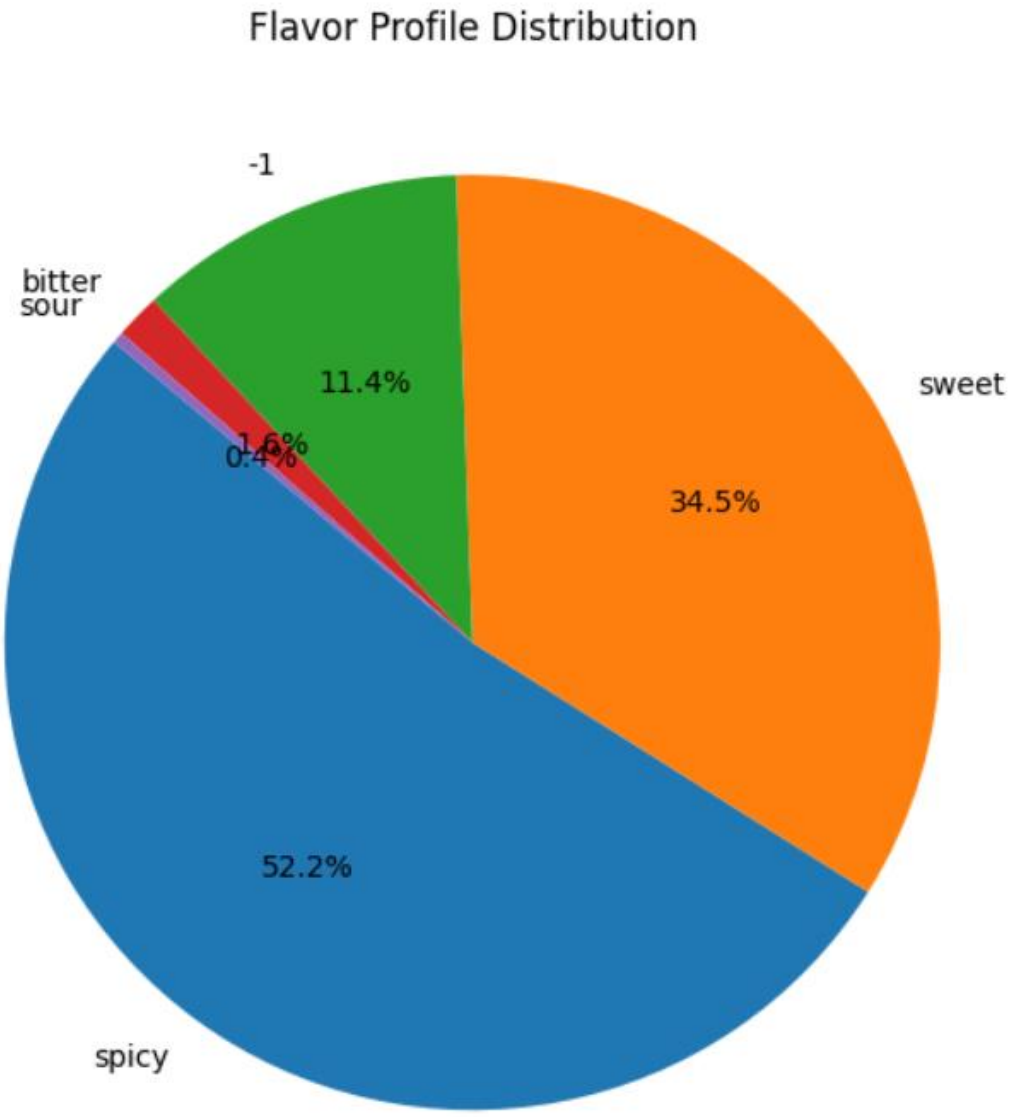
Displays the proportion of dishes across different flavor profiles like sweet, spicy, sour, etc.

Insight:

Spicy dishes dominate the dataset, indicating popular preferences in Indian cuisine.

Importance:

Reveals consumer taste trends and guides menu planning.



10. Heat Map – Correlation Between Numeric Features

Purpose:

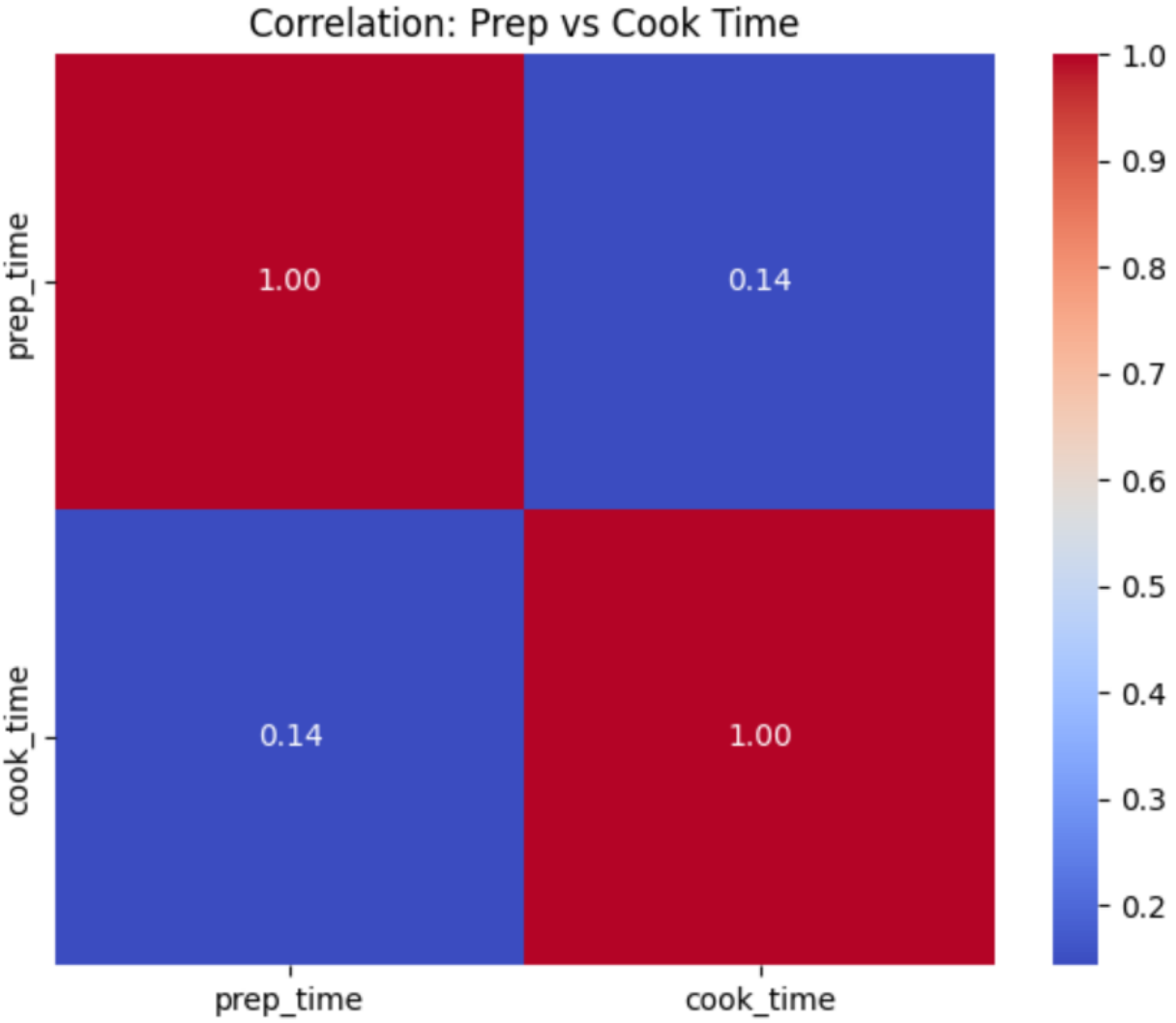
Shows the correlation matrix between numerical features like prep_time, cook_time, etc.

Insight:

Helps identify which numeric fields are related. For example, a moderate positive correlation may exist between prep and cook times.

Importance:

Supports feature selection in modeling and deeper analysis.



9. Findings and Insights

The analysis revealed that:

- Regional Differences: Southern India had a significantly higher number of dishes in the dataset compared to other regions.
- Cooking Times: Most dishes took less than 45 minutes to cook, but certain regional specialties took longer.
- Ingredient Usage: The most common ingredients across dishes were onions, tomatoes, and garlic.
- Flavor Trends: Spicy dishes dominated the dataset, particularly in Southern and Eastern India.

10.Conclusion

This project has successfully demonstrated the power of data science techniques in uncovering insights from food restaurant data. By performing a comprehensive analysis of the dataset, we were able to extract valuable patterns and trends related to regional preferences, ingredient usage, and cooking times, thereby gaining a deeper understanding of the dynamics within the food industry.

The exploratory data analysis (EDA) and various visualizations such as word clouds, line plots, bar charts, and scatter plots provided a clear picture of the underlying relationships in the data. The insights gained from these visualizations are not only useful for understanding customer preferences but also offer actionable information for restaurant owners and food industry stakeholders to enhance their offerings.

The key findings from this project include:

- Ingredient Usage: Identifying the most common ingredients used in dishes provided insights into the staples of Indian cuisine, highlighting the diversity and variety of food choices.
- Regional Preferences: The analysis revealed significant regional differences in food preferences and dish distribution, helping understand the culinary diversity across India.
- Cooking and Preparation Times: The relationship between prep time and cooking time was examined, providing a clearer understanding of how the complexity of dishes influences their preparation and cooking durations.
- Flavor Profiles: The flavor profile analysis highlighted the prominence of spicy dishes in the dataset, reflecting the traditional flavor preferences in Indian cuisine.

Through this project, we have successfully demonstrated how data science and analytics can be applied to the food industry, allowing businesses to make more informed decisions and better cater to consumer preferences. The insights derived from this analysis can be further explored to drive personalized recommendations, optimize menu offerings, and improve overall operational efficiency.

11. Future Scope

Future work can include:

- Incorporating User Feedback: Integrating user reviews and ratings for dishes.
- Personalized Recommendations: Using machine learning to suggest dishes based on user preferences.
- Real-time Data: Incorporating real-time restaurant data for more dynamic analysis.

12. References

- Kaggle. (2025). *Food Restaurant Dataset*. Retrieved from Kaggle Website.
- VanderPlas, J. (2016). *Python Data Science Handbook*. O'Reilly Media.
- Matplotlib Documentation: https://matplotlib.org
- Seaborn Documentation: https://seaborn.pydata.org