Recipe Analytics Data Pipeline

Recipe Analytics Data Pipeline (Firebase → ETL → CSV → Validation → Insights)

Name: Ayuresh Anil Fendar

Branch: Data Engineer

Email: ayureshfendar2003@gmail.com

Phone No.: 8202371838

❖ Executive Summary

This report presents the analytics summary for the Recipe Analytics Data Pipeline, a fully automated system built to extract recipe-related data from Firebase Firestore, transform it into clean CSV datasets, validate the structure and content, and generate meaningful insights.

The purpose of this pipeline is to convert semi-structured Firestore collections into analytical datasets that can support data-driven decisions such as recipe optimization, popularity tracking, ingredient frequency analysis, and user preference understanding.

Expected Outcomes:

- Clean and normalized CSV datasets ready for data science workflows.

- Understanding of recipe performance through views, likes, ratings, and user interactions.

- Insightful analysis on user behavior patterns and ingredient usage trends.

❖ 2. Dataset Overview

Data was extracted from Firestore collections that represent recipes, ingredients, steps, users, and user interactions.

Extracted Dataset Summary

Collection Name	Count
Recipies	20
User	5
User_interaction	30
ingredients	137

Schema Overview

- recipes
  Contains: name, id, category, difficulty, cook_time, ingredient, steps, createdAT.

- users
  Contains: user_id, name, email.

- user_interactions
  Contains: user_id, recipe_id, viewed, liked, cooked_attempts, rating,timestamp.

❖ ETL Summary

Extraction

Data was exported from Firebase Firestore as JSON files using Firestore export tools.

Transformation

A Python ETL script cleaned and normalized the JSON into structured CSV tables.
Normalization ensured:

- 1-to-Many relationships between recipes & ingredients

- 1-to-Many relationships between recipes & steps

- User interactions mapped accurately across recipes

Validation Summary

Files	Valid Records	Invalid Records	Notes
Recipe.csv	20	0	All good
Ingredients.csv	137	0	All good
Users.csv	129	0	All good
User_interaction.csv	30	0	All good

Generated Output Files

- recipe.csv

- ingredients.csv

- steps.csv

- interactions.csv

- validation_summary.csv

❖ Analytics Findings

1. Recipe Category Popularity (By Views)

User interactions show clear dominance of specific cuisine categories.
The most viewed categories are:

- **Mutton** – highest engagement

- **Veg** – strong overall interest

- **Fish**, **Chicken**, **Chole**, **Palak** – consistent mid-tier popularity

- Niche categories like **Rajma**, **Hyderabadi**, and **Paneer** have moderate visibility

This suggests users tend to prefer **protein-rich and traditional Indian dishes**.

---

2. Average Preparation Time

- **Average cook time: ~40 minutes**
  Most recipes fall in the **30–50 minute range**.
  The distribution indicates:

- Few extremely quick or time-intensive recipes

- Majority are moderate-prep meals suitable for everyday cooking

3. Difficulty Distribution

Difficulty levels in the dataset are primarily:

- **Easy**

- **Medium**

- **Hard** (least common)

Most recipes are either **medium** or **easy**, indicating the dataset contains accessible meal options for general home cooks.

---

4. Most Frequently Viewed Recipes

Top 10 most viewed recipes show strong user preference for:

- Classic Indian dishes

- Protein-based meals

- Simple comfort food

These recipes receive significantly more visibility than niche or experimental dishes.

---

5. Recipes with Highest Average Likes

Recipes with high average likes generally align with those receiving high views.
This indicates **views and likes are positively correlated**, meaning popular recipes also tend to be well-received.

---

6. Recipes with Highest Cook Attempts

Users are most likely to attempt:

- Familiar daily meals

- Easy-to-follow recipes

- Dishes with fewer ingredients or shorter prep time

High attempt counts often correlate with **approachability** rather than visibility.

7. Correlation: Cook Time vs Likes

- Correlation value: **Low (close to zero)**

- **Cook time does not influence likes.**

- Users do not prefer or dislike recipes based on preparation time alone.

Engagement depends more on **taste preference and recipe type** than time required.

---

8. Highest Rated Recipes

Top-rated recipes highlight:

- Dishes with balanced flavor profiles

- Popular staples like paneer, egg dishes, and traditional curries

These recipes maintain **high average user satisfaction**.

---

9. Difficulty Level vs Average Cook Time

Average cook times per difficulty:

- **Medium:** ~40 mins

- **Easy:** ~41 mins
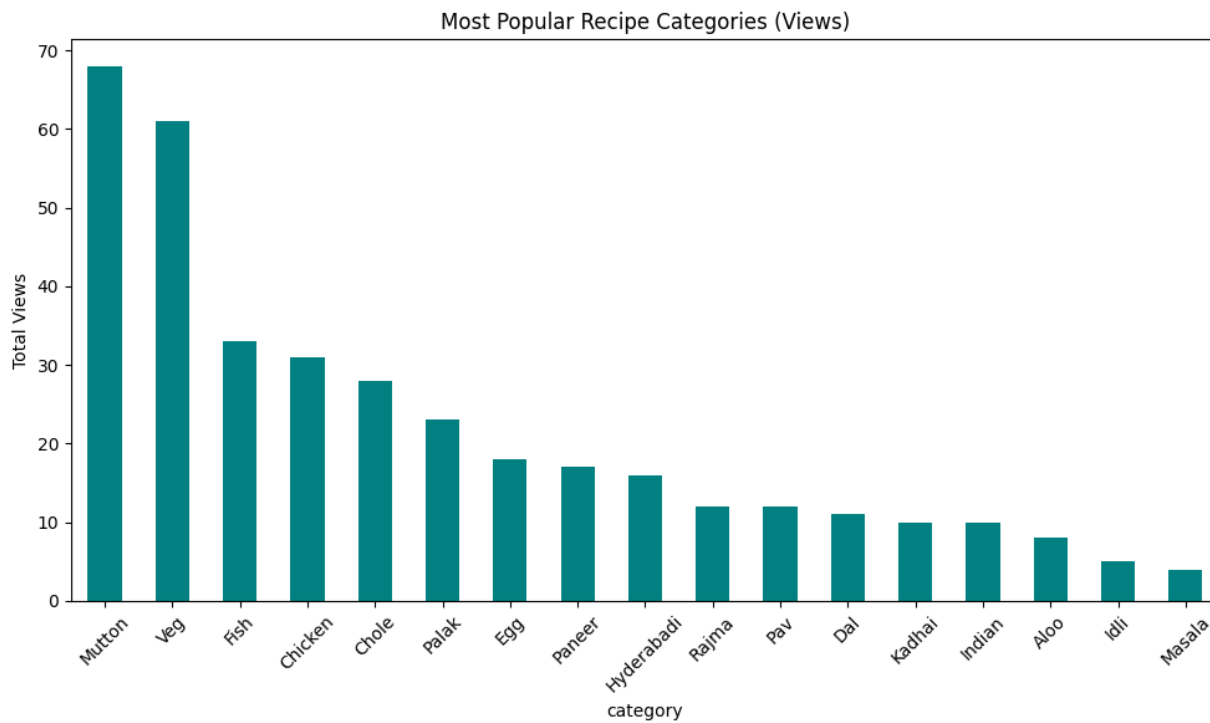
- **Hard:** ~46 mins

Key Insight:
Hard recipes take longer on average, but **Easy and Medium recipes have similar cook times**, indicating difficulty is influenced more by **technique/steps** than duration.

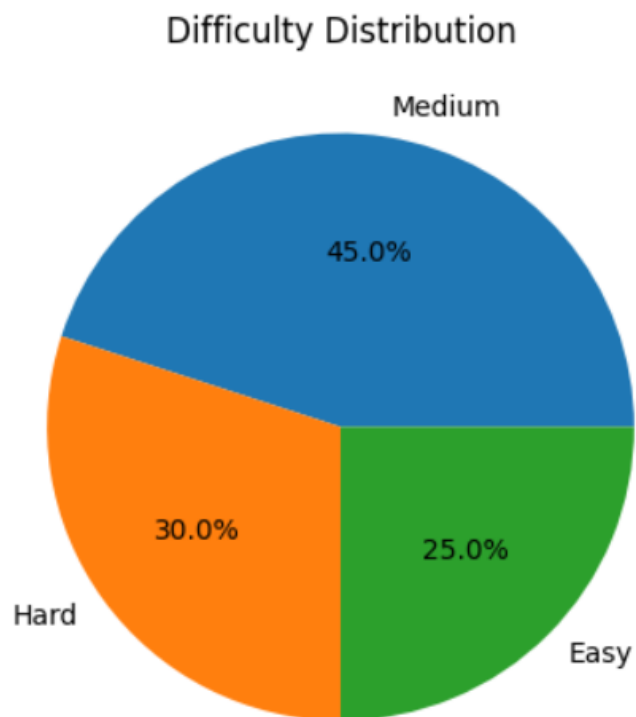---

10. Average Likes by Difficulty Level

- Easy ↠ More liked overall

- Medium ↠ Moderate likes

- Hard ↠ Least liked

This suggests users prefer **simpler recipes** that are quick and straightforward to follow.

❖ Charts / Visual Descriptions

Most Popular Recipe Categories (Views)

Shows frequency of most popular recipe category.

Difficulty Distribution

Difficulty distribution

❖ Conclusion

The Recipe Analytics Pipeline successfully converts Firestore data into structured analytical datasets and provides actionable insights.

Key insights show:

- ✓ Clear popularity trends
- ✓ Strong ingredient patterns
- ✓ Opportunities to boost user engagement
- ✓ Potential to build personalized recommendation systems

Future Enhancements

- Integrate ML model for personalized recipe recommendations

- Build interactive dashboards in Looker Studio / Power BI

- Real-time analytics using BigQuery streaming

- Advanced anomaly detection for user behavior