Exploratory Data Analysis - report

Maciej S.

October 2024

1 Introduction

1.1 What is EDA?

EDA stands for Exploratory Data Analysis. It's a crucial step in the data analysis process where you summarize the main characteristics of a dataset, often using visual methods. Here's the gist:

- Purpose: Understand the data better before jumping into modeling. It helps in uncovering patterns, spotting anomalies, testing hypotheses, and checking assumptions.
- Tools and Techniques: Descriptive statistics (mean, median, mode, standard deviation), visualizations (histograms, scatter plots, box plots), and data wrangling (handling missing values, scaling).
- Outcome: You get insights that guide your next steps in the data science workflow.

You can think of it as getting to know your dataset inside out, making sure you're fully prepped before the heavy lifting begins.

1.2 The aim

This analysis aims to understand the dataset of 134 cocktails. It aims to identify main characteristics of the cocktails, understand relations between ingredients and distinguish potential groups of similar drinks.

1.2.1 Brief overlook of the dataset

Main part of the data collection consists of **134** rows and **11** columns. Here is brief look at it:

id	name	category	glass	tags	instructions	imageUrl	alcoholic	createdAt	updatedAt	ingredients
11000	Mojito	Cocktail	Highball glass	[IBA, ContemporaryClassic, Alcoholic, USA, Asi...	Muddle mint leaves with sugar and lime juice. ...	https://cocktails.solvro.pl/images/ingredients...	1	2024-08-18T19:01:17.000+00:00	2024-08-18T19:06:16.000+00:00	[{'id': 170, 'name': 'Soda water', 'descriptio...
11001	Old Fashioned	Cocktail	Old-fashioned glass	[IBA, Classic, Alcoholic, Expensive, Savory]	Place sugar cube in old fashioned glass and sa...	https://cocktails.solvro.pl/images/ingredients...	1	2024-08-18T19:01:58.000+00:00	2024-08-18T19:06:17.000+00:00	[{'id': 513, 'name': 'Water', 'description': '...

Figure 1: TheCocktailDB head

The columns are: id, name, category, glass, tags, instructions, imageUrl, alcoholic, createdAt, updatedAt, ingredients. Every column excpet 'id' and 'alcoholic' have object type. Other two are represented as int64.

In 'ingredients' column there is nested information about ingredients. Each drink have own piece of ingredients information there. If you unpack such block of information there will be data about ingredients needed to prepare such drink. Every record in 'ingredients' table has 10 attributes. They are as follows: id, name, alcohol, type, percentage, imageUrl, createdAt, updatedAt, measure. After unpacking ingredients data from every cocktail there will be **531** unique records in the 'ingredients' table.

id	name	description	alcohol	type	percentage	imageUrl	createdAt	updatedAt	measure
170	Soda water	None	1	None	NaN	None	2024-08-18T19:01:57.000+00:00	2024-08-18T19:01:57.000+00:00	NaN
305	Light Rum	Light rums, also referred to as "silver" or "w...	1	Rum	NaN	https://cocktails.solvro.pl/images/ingredients...	2024-08-18T19:02:37.000+00:00	2024-08-18T19:02:37.000+00:00	2-3 oz
312	Lime	A lime (from French lime, from Arabic lima, fr...	0	Fruit	NaN	https://cocktails.solvro.pl/images/ingredients...	2024-08-18T19:02:40.000+00:00	2024-08-18T19:02:40.000+00:00	Juice of 1
337	Mint	Lamiaceae (/ˌleɪmiˈeɪsiˌaɪ/ or /ˌleɪmiˈeɪʃiː/...	0	Flower	NaN	https://cocktails.solvro.pl/images/ingredients...	2024-08-18T19:02:47.000+00:00	2024-08-18T19:02:47.000+00:00	2-4
476	Sugar	Sugar is the generic name for sweet-tasting, s...	0	None	NaN	https://cocktails.solvro.pl/images/ingredients...	2024-08-18T19:03:31.000+00:00	2024-08-18T19:03:31.000+00:00	2 tsp

Figure 2: A snippet of the 'ingredients' table

2 Data review

...