

# Steps for Zomato data wrangling:

## Data extraction:

- The data is extracted from Zomato website using Python package '**Beautiful Soup**'.
- This is data is for all restaurants of Bangalore city which is around 12k-13k of records, pulled as of January 2020.
- Following are the fields
  1. **restaurant\_link**: Link for the restaurant
  2. **restaurant\_ID**: Unique restaurant id
  3. **restaurant\_name**: Name of the restaurant
  4. **locality**: neighbourhood of the restaurant
  5. **restaurant\_category**: Category of restaurant based on what food they serve, like dining or quick bites, etc.
  6. **zomato\_gold**: Whether the restaurant provides zomato gold benefits
  7. **discounts**: Discounts offered by the restaurant
  8. **photos\_taken**: Number of photos taken at the restaurant
  9. **rating**: Zomato rating
  10. **votes**: Votes for the ratings or reviews
  11. **cuisines**: Type of cuisines served
  12. **approx.\_cost\_for\_2**: Approx cost for 2 people
  13. **opening timings**: Opening and closing timings of the restaurant
  14. **address**: Detailed address of the restaurant
  15. **latitude**: Latitude of restaurant
  16. **longitude** :Longitude of the restaurant
  17. **more\_info**: main features or services provided by the restaurant like delivery, outside seating, etc
  18. **featured\_in**: Featured in which categories of Zomato collections
  19. **most\_liked\_food**: Most liked or famous for in food items and rating

20. **most\_liked\_service**: Most liked service of the restaurant and rating
21. **most\_liked\_look&field**: Most liked, look and feel of the restaurant and rating
22. **reviews**: Reviews available on first page of the restaurant along with time of review posted and sentiments.

### **Data cleaning:**

- Most of the data is cleaned/formatted while scraping.
- Some columns are manipulated to tuples.
- Opening and closing timings are transformed to datetime formats.
- Missing values are transformed to np.NAN
- Duplicates rows, if any, are removed based on the restaurant\_id.
- No outliers.

For Data cleaning and EDA refer:

[https://github.com/Anandpatil412/Zomato\\_Res\\_analysis\\_Pred\\_Rating/blob/main/DataWrangling/zomatoDataCleaning.ipynb](https://github.com/Anandpatil412/Zomato_Res_analysis_Pred_Rating/blob/main/DataWrangling/zomatoDataCleaning.ipynb)

For Web scarping refer:

[https://github.com/Anandpatil412/Zomato\\_Res\\_analysis\\_Pred\\_Rating/blob/main/DataExtraction\(WebZomato\)/zomatoScrapper.ipynb](https://github.com/Anandpatil412/Zomato_Res_analysis_Pred_Rating/blob/main/DataExtraction(WebZomato)/zomatoScrapper.ipynb)