ASSIGNMENT SUBMISSION FORM

This will be the first page of your assignment

Course Name: AMPBA
Assignment Title: DCPP group assignment
Submitted by: DCPP group 17

(Student name or group names)

Student Name	PG ID
Kenny Devarapalli	12120062
Mrinal Chitranshu	12120016
Nagraj G T	12120095
Hari Teja	12120029
Madhab Chakraborty	12120045

EXECUTIVE SUMMARY

Problem statement:

There is plenty of information available in the websites regarding different food recipes. Most of them are related to the preparation process, ingredients, time requirement for preparation, also there are some nutritional information available which could be helpful in different medical conditions for selection of different recipes of various categories.

However, getting a single resource from where we can get all intuitive information (with data support) is challenging. In fact, it would be more helpful if we can get some information as mentioned below;

- Healthy Dishes
- Gluten free dishes
- Quick preparation dishes
- Dishes with less number of ingredients
- Recipes to avoid for diabetic patients,
- Recipes to avoid for cardiac disease patients,
- Recipes to avoid for chronic kidney diseases patients.

Proposed solution:

We have accumulated all data in one place from where we can able to get information which can address the above-mentioned queries. Also, we want to find out the relations like preparation time vs nutritional value, etc. The same will also address popular recipes for the persons with various health issues.

Challenges:

The challenges faced:

- Right resource
- Permission for data scraping from the right resources
- Availability of all relevant data
- Null values in the data source
- Selection of right data from the webpage
- Duplicity of recipes

THE CHOSEN DOMAIN AND SEED SOURCES

Reason behind choosing "Food and Recipe" domain:

- Food is one of the "basic and essential need".
- So, it is important to choose the right food as well. The choice of food again depends on various factors like time of intake, level of hunger, geographical location, preferential type (veg/non-veg etc), health/medical condition, even on the socio-economic status etc.
- Unfortunately, we give little thoughts to all these factors while choosing the recipes but gives top priority to the taste and predefined mental setup. The time is changing and we have also started to give priority to "Healthy food" as well.

Food is one of the "basic need" and the choice of right food/recipe is also important. Moreover, as it has a large variety and diversity, we found ample opportunity to work in this field which driven us towards the selection of this domain.

We have collected the data related to Indian food recipes from different websites. The selection of the sources was made based on the reach of the website, authenticity, popularity, and ratings. We have considered availability of various data in the chosen websites relevant with variety of foods, ingredients, different categories (breakfast, main course, veg, non-veg, dessert, juices, healthy food). Also, we have given a special emphasis on the nutritional values and components to facilitate all for selection of recipes in different medical restrictions.

THE STRUCTURED AND UNSTRUCTURED SOURCES FROM OPEN DOMAIN/ INTERNAL SOURCES

Sources (open domain):

1. https://www.veganricha.com/
2. https://www.allrecipes.com/

Sources (internal sources):

Given source in ISB LMS

3. We have chosen https://www.veganricha.com/ and https://www.allrecipes.com/ because these sites have good collection of recipes and mostly whatever we can think of Indian food recipes are present. There were very few absent fields present and data in html was also in a well-formatted form. Automation through selenium for clicks while scraping the data was also possible in these, provided we follow the policies published by them.

   Personally, one of our team members was using these websites for recipe reference, as he is connoisseur of food and drinks and so he recommended these websites. Even we all were also amazed to see that most of the Indian recipes were present and they also had many foods what even we have never heard of, but definitely want to try.

DOWNLOAD/ CRAWL/ COLLECT DATA FROM ALL THE SOURCES METHODS USE:

Programming: Python

Python packages

1. BeautifulSoup
2. Request
3. selenium
4. sweetviz
5. time
6. urllib.request
7. workcloud

For automation:

1. Selenium
2. Webdriver

CHALLENGES AND RESOLVE:

1. **Block of our IP address by the url:**
    - We faced block of our IP address by the url while we perform web scrapping/crawling operation. We overcome the situation by using "time" package by making the system to forced sleep for few seconds.
    - Sometimes, we used VPN and changed the location.
    - Distributed tasks among team members so that frequent hit/ downloading of data won't happen from a single IP.
2. **Null values:**
    - We have overcome the issue by delete that specific row
3. Installing workcloud in mac. It was bit difficult but we managed to do it taking help from geekforgeeks and stackOverflow
4. Installing webdriver in mac. It was bit difficult but we managed to do it taking help from geekforgeeks and stackOverflow

CONVERT DATA FROM ORIGINAL SOURCES (WEBPAGES, PDF FILES, CSV FILES) TO STRUCTURED DATA FIELDS

We used python programming and various python packages to convert data into expected format. Packages used were pandas, numpy, BeautifulSoup and request. For data collected from various sources, the main challenge was to merge the data into a common dataframe and to decide the columns. This is because different websites had different names for the xml/html tags containing similar types of values. E.g- In one source, all ingredients were in a single cell but in another source, each ingredient was present in a separate cell.

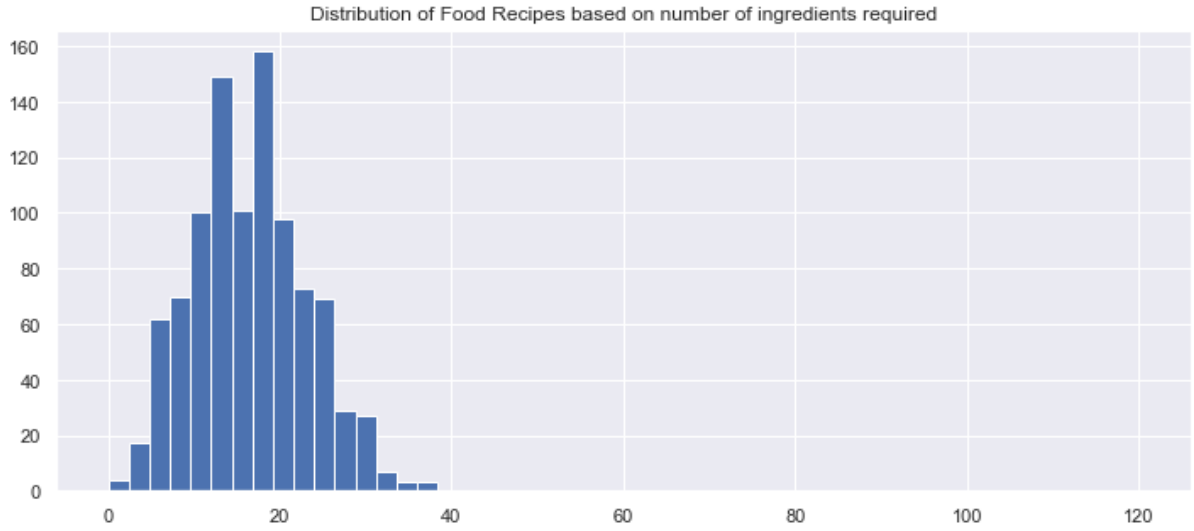
Number of fields also had a wide difference in various scraped sources.

DATA CLEANING/PRE-PROCESSING

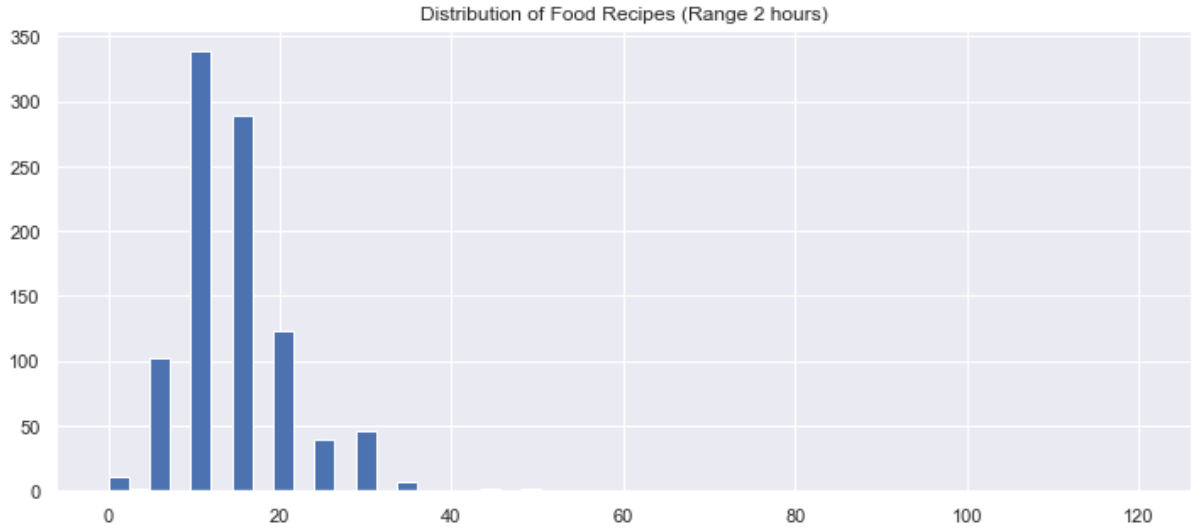
Techniques

- Dropping marginal no. of rows in a DataFrame, having absent values.
- Conversion of column to common data type.
- Skipping rows having duplicate entries, collected from various sources.
- Replacing null values with zeros in limited no. of rows in order to maintain uniformity of data type.
- Renaming columns to make column names more readable
- Appending data collected from multiple sources by choosing same columns from both datasouces.

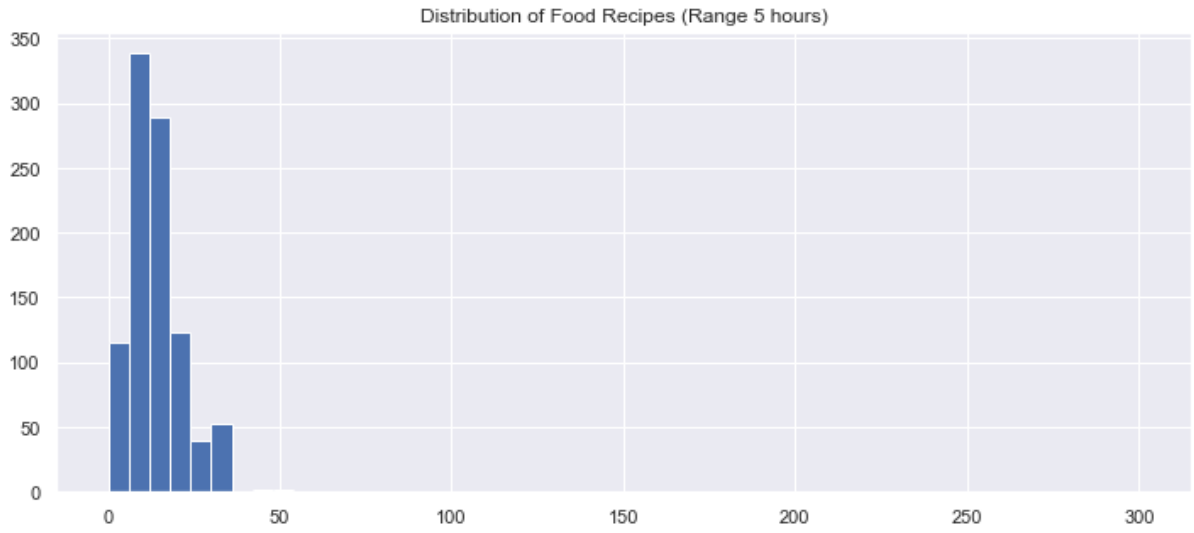
OBSERVATIONS/ INSIGHTS AND ANALYSIS ON THE DATA COLLECTED

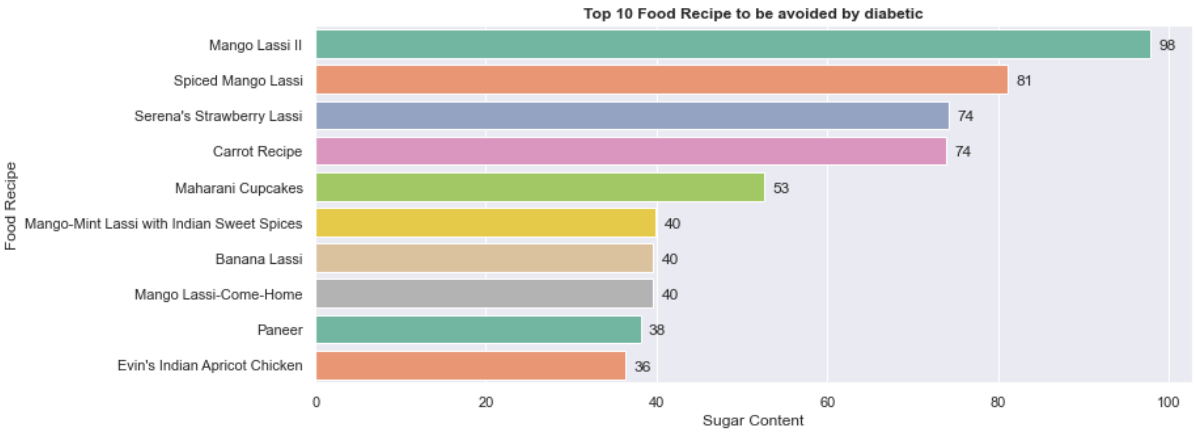


Above graph represents the number of foods recipes based on the number of ingredients required. We could see that most of the food items requires 12-18 ingredients. Also can observe the data is somehow normally distributed.

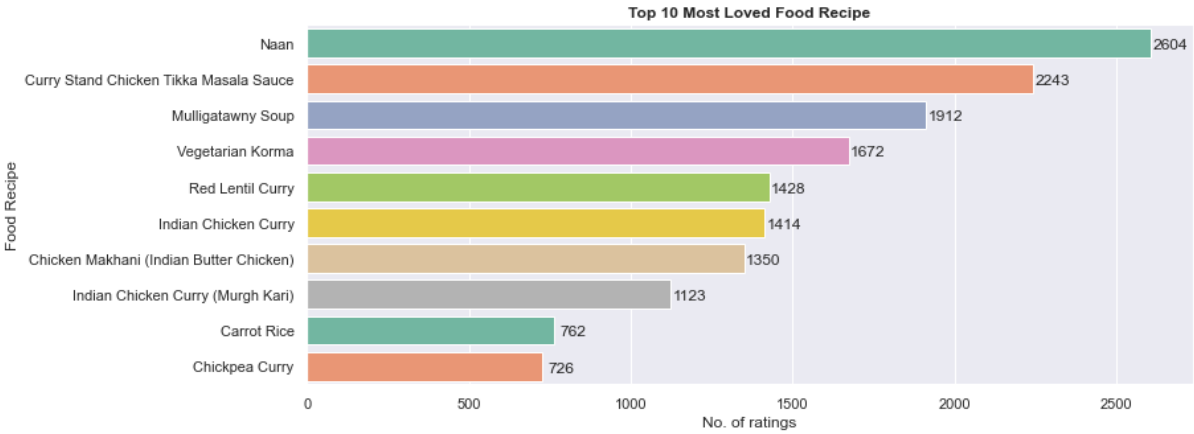


Above graph represents the number of foods recipes based on the preparation time in minutes required. On an average, based on the data most of the food items can be prepared in 10-15 minutes. The data is distributed normally with bit rightward skewness



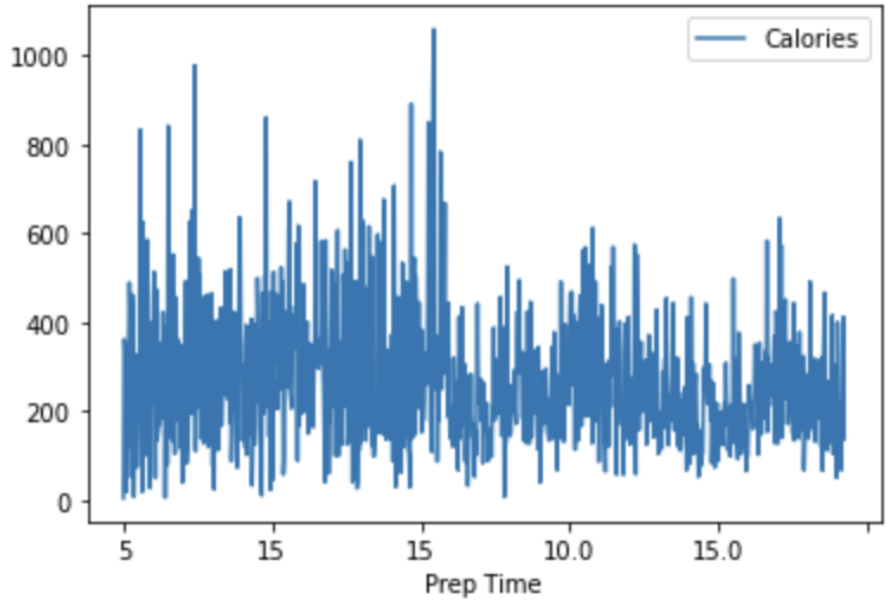


Above graph represents the top 10 food items to be avoided by patients having diabetes as comorbid disease.



Above graph represents the top 10 most loved food recipes based on rating given by connoisseurs.

```
data.plot(x = 'Prep Time',y = 'Calories', kind='line')
plt.show()
```



Above graph represents the relationship between calories content and respective preparation time. No much significance can be observed between calories and preparation time based on the collected data.

Most Used Keywords

STRATEGY TO ENHANCE THE DATA WITH CROWD SOURCING METHODS

There are a number of strategies to enhance Indian recipe data using crowd sourcing.

1. Scraping Instagram accounts of food influencers/bloggers to see which recipe is being more liked, viewed and talked about in the comment
2. Scraping YouTube influences making Indian recipes, especially Indian cooking shows. The likes and view count tells us how each recipe is appreciated or not.
3. Flood blogging websites allow bloggers to write about their favourite food recipes. Followers of the blog react to the same.
4. Hospital and Doctor prescriptions on which food to eat and not to eat based on disease or not to can be incorporated into the Recipe data.
5. Website such as Zomato and Swiggy allow users to buy whichever food item they want. We can see the frequency of which food item is purchased more and at what average price.

REFERENCES AND SOURCES USED FOR THIS ASSIGNMENT

Sources (open domain):

1. https://www.veganricha.com/
2. https://www.allrecipes.com/

Sources (internal sources):

Given source in ISB LMS

Technical Reference:

https://stackoverflow.com/
https://www.geeksforgeeks.org/