WEB SCRAPING SIMPLY RECIPES

KATRYN | CAITLIN
HEESUNG | WILLIAM

ON TODAY'S MENU

01

WELCOME

OUR TEAM

02

PROBLEM

POP-UPS SUCK

03

SOLUTION

BEAUTIFULSOUP GAVE US A BEAUTIFUL IDEA

04

SCRAPING & DATA

DATA ALPHABET SOUP

05

CHARTS

LETS VISUALIZE RECIPES WITH PLOTTING

06

FLASK

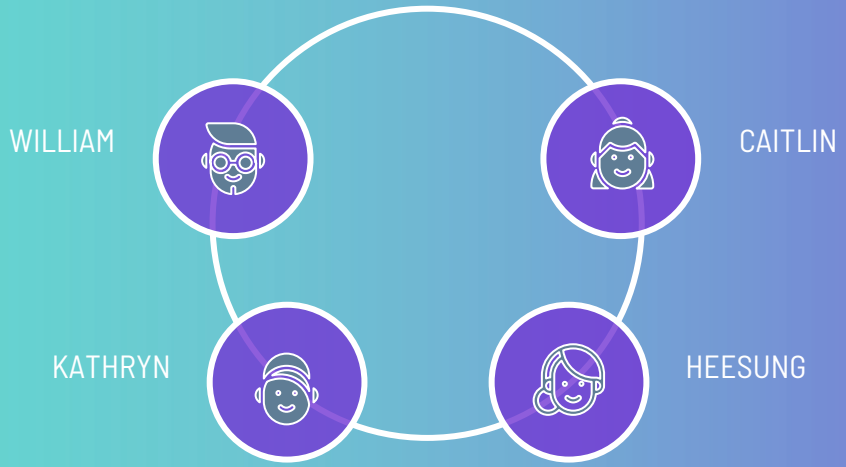
HOW DO WE PRESENT OUR FINDINGS

01 WELCOME

THE REBOOT IS REAL...

THE CIRCLE OF LIFE

Despite our best efforts, we
couldn't live without each other



OUR PROBLEM

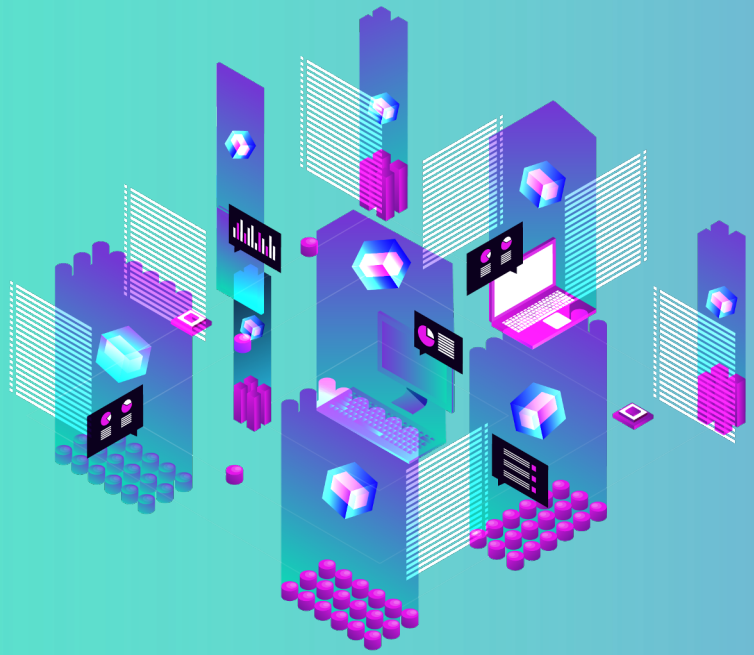
02

Pop-ups suck!

POP-UPS SUCK, PERIOD

The team visited countless recipe and food-related websites, hoping to find an example of a recipe website without BS. Instead we found…

- CLICK POP-UPS
- TIMED POP-UPS
- SCROLL POP-UPS



SOLUTION 03

BEAUTIFULSOUP GAVE US A
BEAUTIFUL IDEA

GUIDING PRINCIPLES

MISSION

To create a simpler way for a user to navigate recipes on the popular recipe website 'SimplyRecipes.com', by automating the process

VISION

Harness the power of BeautifulSoup & Selenium to do ETL on SimplyRecipes data –

1. Scraping the data
2. Transforming the data
3. Loading into Mongo DD (Atlas Cluster)

STRATEGY

STEP 1 - EXTRACTION

By harnessing the power of Python and libraries such as Selenium and BeautifulSoup, SimplyRecipes was scraped for each recipe's key attributes.

STEP 3 - LOADING

We eventually had a JSON file. We decided to upload to a Mongo DB Atlas Cluster, for complete collaborativeness

STEP 5 – ANALAYSIS & VISUALIZATION

We created a RESTFUL API through a Python Flask-powered app using JavaScript, CSS, HTML & Bootstrap

STEP 2 - TRANSFORMATION

The output text file from our scrape required transformation, before in terms of data quality and integrity.

STEP 4 - PLOTTING

By using Jupyter Lab in conjuction with various Python libraries, we set about plotting the data from eahc scraped recipe

```
driver=webdriver.Chrome('chromedriver.exe')
url = "https://www.simplyrecipes.com/index/"
driver.get(url)
response=requests.get(url)
soup=BeautifulSoup(response.text,'html.parser')
driver.maximize_window()

recipe_list=[]
for link in linklist_text[1 :]:
    time.sleep(0.3)
    target=driver.find_element_by_partial_link_text(link)
    target.click()
    time.sleep(0.1)
    cards = driver.find_elements_by_class_name("grd-title-link")
    for i in range(0,len(cards)):
        try:
            newcards = driver.find_elements_by_class_name("grd-title-link")
            time.sleep(0.3)
            newcards[i].click()
            time.sleep(0.3)
            recipe=driver.find_element_by_id("sr-recipe-callout")
            recipe_list.append(recipe.text)
            driver.back()
            time.sleep(0.3)
        except:
            continue
    nxt=driver.find_elements_by_class_name("rpg-next")
    if len(nxt) > 0:
        pages=driver.find_elements_by_class_name("rpg-page-numbers")
        textpages=pages[-2].text
        lastpage=int(textpages)
        for i in range(1,lastpage):
            nxt2=driver.find_elements_by_class_name("rpg-next")
            time.sleep(0.3)
            try:
                nxt2[0].click()
                cards2 = driver.find_elements_by_class_name("grd-title-link")
                for i in range(0,len(cards2)):
                    try:
                        newcards2 = driver.find_elements_by_class_name("grd-title-link")
                        time.sleep(0.3)
                        newcards2[i].click()
                        time.sleep(0.3)
                        recipe=driver.find_element_by_id("sr-recipe-callout")
                        recipe_list.append(recipe.text)
                        driver.back()
                        time.sleep(0.3)
                    except:
                        continue
            except:
                continue
    with open ("recipe_list2.txt", "a") as fout:
        for recipe_text in recipe_list:
            fout.write("%s\n" % recipe_text.encode("utf-8"))
            recipe_list=[]
    driver.get(url)
```

LET'S TALK CODE

A FEW WORDS FROM KATHRYN

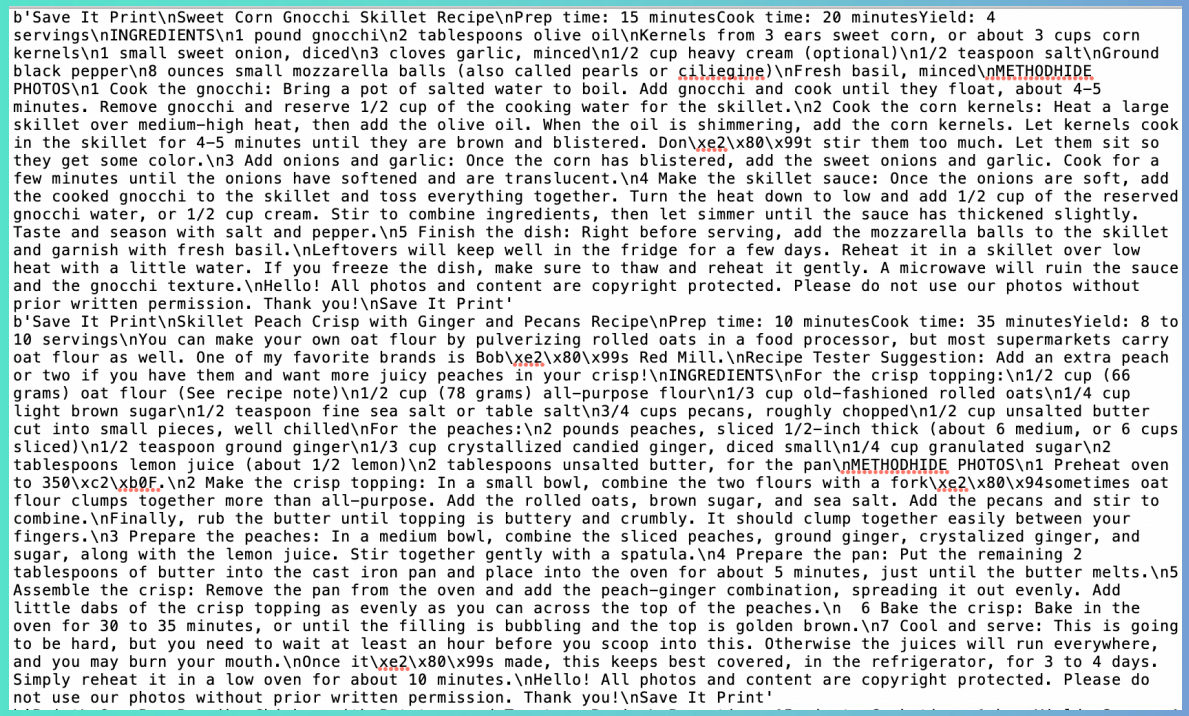
05 CHARTING

Talk charting to me...

Heesung’s chart

A FEW WORDS FROM Heesung

DATA EXTRACTION



```
b'Save It Print\nSweet Corn Gnocchi Skillet Recipe\nPrep time: 15 minutesCook time: 20 minutesYield: 4
servings\nINGREDIENTS\n1 pound gnocchi\n2 tablespoons olive oil\nKernels from 3 ears sweet corn, or about 3 cups corn
kernels\n1 small sweet onion, diced\n3 cloves garlic, minced\n1/2 cup heavy cream (optional)\n1/2 teaspoon salt\nGround
black pepper\n8 ounces small mozzarella balls (also called pearls or ciliegine)\nFresh basil, minced\nMETHODHIDE
PHOTOS\n1 Cook the gnocchi: Bring a pot of salted water to boil. Add gnocchi and cook until they float, about 4-5
minutes. Remove gnocchi and reserve 1/2 cup of the cooking water for the skillet.\n2 Cook the corn kernels: Heat a large
skillet over medium-high heat, then add the olive oil. When the oil is shimmering, add the corn kernels. Let kernels cook
in the skillet for 4-5 minutes until they are brown and blistered. Don\xe2\x80\x99t stir them too much. Let them sit so
they get some color.\n3 Add onions and garlic: Once the corn has blistered, add the sweet onions and garlic. Cook for a
few minutes until the onions have softened and are translucent.\n4 Make the skillet sauce: Once the onions are soft, add
the cooked gnocchi to the skillet and toss everything together. Turn the heat down to low and add 1/2 cup of the reserved
gnocchi water, or 1/2 cup cream. Stir to combine ingredients, then let simmer until the sauce has thickened slightly.
Taste and season with salt and pepper.\n5 Finish the dish: Right before serving, add the mozzarella balls to the skillet
and garnish with fresh basil.\nLeftovers will keep well in the fridge for a few days. Reheat it in a skillet over low
heat with a little water. If you freeze the dish, make sure to thaw and reheat it gently. A microwave will ruin the sauce
and the gnocchi texture.\nHello! All photos and content are copyright protected. Please do not use our photos without
prior written permission. Thank you!\nSave It Print'
b'Save It Print\nSkillet Peach Crisp with Ginger and Pecans Recipe\nPrep time: 10 minutesCook time: 35 minutesYield: 8 to
10 servings\nYou can make your own oat flour by pulverizing rolled oats in a food processor, but most supermarkets carry
oat flour as well. One of my favorite brands is Bob\xe2\x80\x99s Red Mill.\nRecipe Tester Suggestion: Add an extra peach
or two if you have them and want more juicy peaches in your crisp!\nINGREDIENTS\nFor the crisp topping:\n1/2 cup (66
grams) oat flour (See recipe note)\n1/2 cup (78 grams) all-purpose flour\n1/3 cup old-fashioned rolled oats\n1/4 cup
light brown sugar\n1/2 teaspoon fine sea salt or table salt\n3/4 cups pecans, roughly chopped\n1/2 cup unsalted butter
cut into small pieces, well chilled\nFor the peaches:\n2 pounds peaches, sliced 1/2-inch thick (about 6 medium, or 6 cups
sliced)\n1/2 teaspoon ground ginger\n1/3 cup crystallized candied ginger, diced small\n1/4 cup granulated sugar\n2
tablespoons lemon juice (about 1/2 lemon)\n2 tablespoons unsalted butter, for the pan\nMETHODHIDE PHOTOS\n1 Preheat oven
to 350\xc2\xb0F.\n2 Make the crisp topping: In a small bowl, combine the two flours with a fork\xe2\x80\x94sometimes oat
flour clumps together more than all-purpose. Add the rolled oats, brown sugar, and sea salt. Add the pecans and stir to
combine.\nFinally, rub the butter until topping is buttery and crumbly. It should clump together easily between your
fingers.\n3 Prepare the peaches: In a medium bowl, combine the sliced peaches, ground ginger, crystalized ginger, and
sugar, along with the lemon juice. Stir together gently with a spatula.\n4 Prepare the pan: Put the remaining 2
tablespoons of butter into the cast iron pan and place into the oven for about 5 minutes, just until the butter melts.\n5
Assemble the crisp: Remove the pan from the oven and add the peach-ginger combination, spreading it out evenly. Add
little dabs of the crisp topping as evenly as you can across the top of the peaches.\n  6 Bake the crisp: Bake in the
oven for 30 to 35 minutes, or until the filling is bubbling and the top is golden brown.\n7 Cool and serve: This is going
to be hard, but you need to wait at least an hour before you scoop into this. Otherwise the juices will run everywhere,
and you may burn your mouth.\nOnce it\xe2\x80\x99s made, this keeps best covered, in the refrigerator, for 3 to 4 days.
Simply reheat it in a low oven for about 10 minutes.\nHello! All photos and content are copyright protected. Please do
not use our photos without prior written permission. Thank you!\nSave It Print'
```

4,660 recipes were scraped from simplyrecipe.com

The data we needed:

- Title
- Prep time
- Cook time
- Total time
- Ingredients
- Number of Ingredient
- Method
- Number of Cooking Step

Process

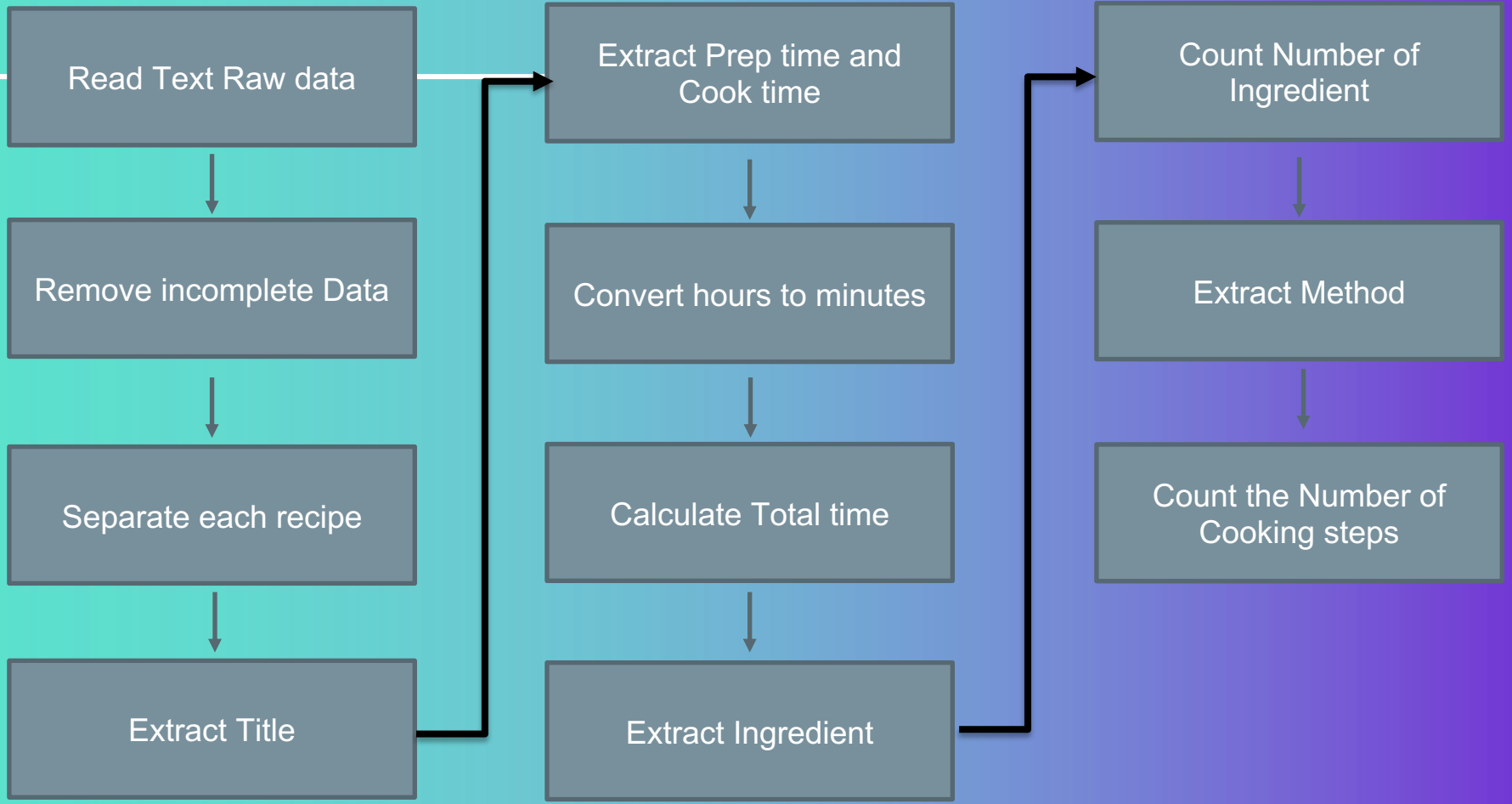
Extract the data – python

Create the dataframe – Pandas

Save the data – CSV and JSON format

Upload the data - MongoDB

EXTRACTING PROCESS



Example Code – Extract Prep time and Cook time

```
['Sweet Corn Gnocchi Skillet Recipe', 'Prep time: 15 minutesCook time: 20 minutesYield: 4 servings', 'INGREDIENTS',
```

```python
title_list=[]
time_line_list=[]
for each_recipe_list in final_recipe_list:
    for line_ind, line in enumerate(each_recipe_list):
        #print(line)
        if "Prep time" in line or "Cook time" in line:
            time_line_list.append(line)
            title_list.append(each_recipe_list[line_ind-1])
#print(title_list)
print(time_line_list)
```

```
['Prep time: 15 minutesCook time: 20 minutesYield: 4 servings', 'Prep time: 10 minutesCook time: 35 minutesYield:
8 to 10 servings', 'Prep time: 30 minutesCook time: 10 minutesYield: 2 servings', 'Prep time: 15 minutesCook time
: 1 hourYield: Serves 4 to 6', 'Prep time: 10 minutesCook time: 15 minutesSteak resting time: 10 minutesYield: Se
rves 4 to 6', 'Prep time: 15 minutesCook time: 35 minutesMarinade time: 2 hoursYield: 4 to 6', 'Prep time: 10 min
utesCook time: 20 minutesYield: Serves 4', 'Prep time: 10 minutesCook time: 20 minutesYield: 6 servings', 'Prep t
ime: 1 hour, 10 minutesCook time: 15 minutesYield: Serves 4', 'Prep time: 12 minutesCook time: 25 minutesYield: 4
```

```python
def get_prep_time_str(a_list,a,b):
    time_str_list = []
    for index,item in enumerate(a_list):
        a_ind = item.find(a)
        b_ind = item.find(b)
        if a_ind == -1 or b_ind == -1:
            time_str_list.append("None")
        else:
            time_str = item[a_ind:b_ind]
            time_str_list.append(time_str)
    print(time_str_list)
    return(time_str_list)

preptime_line_list = get_prep_time_str(time_line_list,"Prep","Cook")
cooktime_line_list = get_prep_time_str(time_line_list,"Cook","Yield")
```

```
['Prep time: 15 minutes', 'Prep time: 10 minutes', 'Prep time: 30 minutes', 'Prep time: 15 minutes', 'Prep time:
10 minutes', 'Prep time: 15 minutes', 'Prep time: 10 minutes', 'Prep time: 10 minutes', 'Prep time: 1 hour, 10 mi
nutes', 'Prep time: 12 minutes', 'Prep time: 10 minutes', 'Prep time: 5 minutes', 'Prep time: 35 minutes', 'Prep
```

CREATE DATAFRAME

```
]:  1 df = pd.DataFrame(list(zip(title_list, prep_time_list,cook_time_list,Total_time_list,ingredient_list,ingredient_
    2                columns =['Name', 'Prep Time','Cook Time','Total Time','Ingredient', 'Number of Ingredient',"Meth
    3 df
```

]:

	Name	Prep Time	Cook Time	Total Time	Ingredient	Number of Ingredient	Method	Number of cooking step
0	Sweet Corn Gnocchi Skillet Recipe	15	20	35	[1 pound gnocchi, 2 tablespoons olive oil, Ker...	10	[1 Cook the gnocchi: Bring a pot of salted wat...	6
1	Skillet Peach Crisp with Ginger and Pecans Recipe	10	35	45	[For the crisp topping:, 1/2 cup (66 grams) oa...	15	[1 Preheat oven to 350\xc2\xb0F., 2 Make the c...	9
2	Steak on the Stovetop Recipe	30	10	40	[One 3/4- to 1-pound steak, 3/4- to 1-inch thi...	5	[1 Pat the steak dry: Pat the steak dry with p...	7
3	One-Pan Paprika Chicken with Potatoes and Toma...	15	60	75	[6 to 8 chicken thighs (about 3 pounds, bone-i...	11	[1 Heat the oven to 400\xc2\xb0F. Place a rack...	8
4	Quick and Easy Pan-Fried Flank Steak Recipe	10	25	35	[1 1/2 pound flank steak, Salt, Freshly ground...	5	[1 Tenderize the steak with shallow cuts: Remo...	15
5	Herbes de Provence Skillet Chicken with Potato...	15	155	170	[For the marinade:, 1 1/2 pounds boneless skin...	19	[1 Label the bag: Use a sharpie to write the d...	14
6	Quick Easy Fish Stew Recipe	10	20	30	[6 tablespoons extra virgin olive oil, 1 mediu...	14	[1 Heat olive oil in a large thick-bottomed po...	3

- Pandas was used.
- The dataframe was created from lists made at the previous extraction.

REMOVE DUPLICATE DATA

```
[ ]: 1 df[df.duplicated(['Name'])==True].sort_values('Name')
```

[ ]:

	Name	Prep Time	Cook Time	Total Time	Ingredient	Number of Ingredient	Method	Number of cooking step
2548	7 Layer Bean Dip Recipe	10	10	20	[2 cups refried beans, from one 15-ounce can o...	12	[1 Prepare the refried beans: Heat the refried...	11
2848	7 Layer Bean Dip Recipe	10	10	20	[2 cups refried beans, from one 15-ounce can o...	12	[1 Prepare the refried beans: Heat the refried...	11
2562	7 Layer Bean Dip Recipe	10	10	20	[2 cups refried beans, from one 15-ounce can o...	12	[1 Prepare the refried beans: Heat the refried...	11
2980	7 Layer Bean Dip Recipe	10	10	20	[2 cups refried beans, from one 15-ounce can o...	12	[1 Prepare the refried beans: Heat the refried...	11
3370	7-Layer Magic Bars Recipe	5	30	35	[2 cups (225 g) graham cracker crumbs, 1/2 cup...	8	[1 Preheat oven to 350F. Spray a 9 x 13-inch b...	9
3460	African Chicken Peanut Stew Recipe	20	115	135	[2-3 pounds chicken legs, thighs and/or wings,...	14	[1 Brown the chicken. Heat the vegetable oil i...	7
2696	African Chicken Peanut Stew Recipe	20	115	135	[2-3 pounds chicken legs, thighs and/or wings,...	14	[1 Brown the chicken. Heat the vegetable oil i...	7

- Out of 3503 recipes, 1945 duplicated recipes were removed.

Save the data as CSV and JSON format

```
]:  1 df.to_csv('final_recipe_dataset.csv')

]:  1 df.to_json(r'final_recipe_dataset.json', orient = "records")

]:  1 final_recipe_dataset_json=df.to_json(orient = "records")
    2 parsed = json.loads(final_recipe_dataset_json)
    3 print(json.dumps(parsed, indent=4, sort_keys=True))
```


```
[
    {
        "Cook Time": 20,
        "Ingredient": [
            "1 pound gnocchi",
            "2 tablespoons olive oil",
            "Kernels from 3 ears sweet corn, or about 3 cups corn kernels",
            "1 small sweet onion, diced",
            "3 cloves garlic, minced",
            "1/2 cup heavy cream (optional)",
            "1/2 teaspoon salt",
            "Ground black pepper",
            "8 ounces small mozzarella balls (also called pearls or ciliegine)",
            "Fresh basil, minced"
        ],
        "Method": [
            "1 Cook the gnocchi: Bring a pot of salted water to boil. Add gnocchi and cook until they float, abou
t 4-5 minutes. Remove gnocchi and reserve 1/2 cup of the cooking water for the skillet.",
            "2 Cook the corn kernels: Heat a large skillet over medium-high heat, then add the olive oil. When th
```

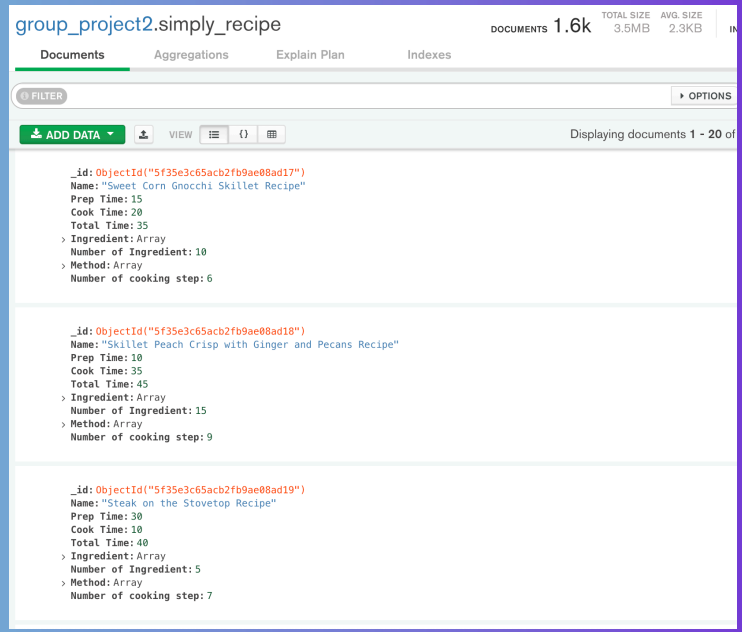

Total 1558 recipes were saved as csv and json format

Upload the data to MongoDB

```
client = MongoClient('localhost', 27017)
db = client['group_project2']
collection_recipe = db['simply_recipe']
```

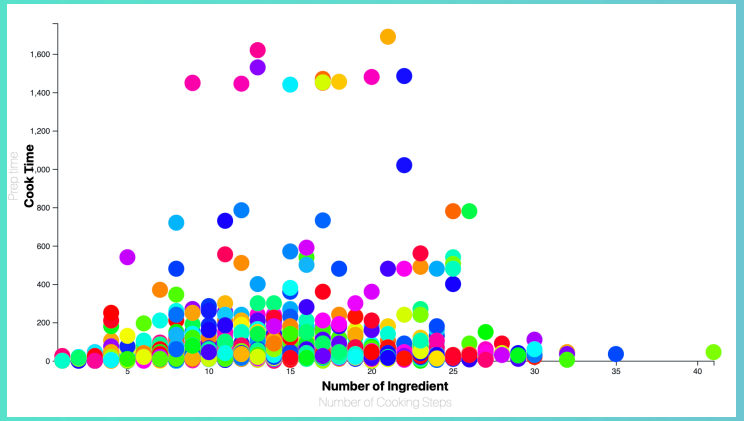
```
with open('final_recipe_dataset.json') as f:
    file_data = json.load(f)
collection_recipe.insert_many(file_data)
client.close()
```

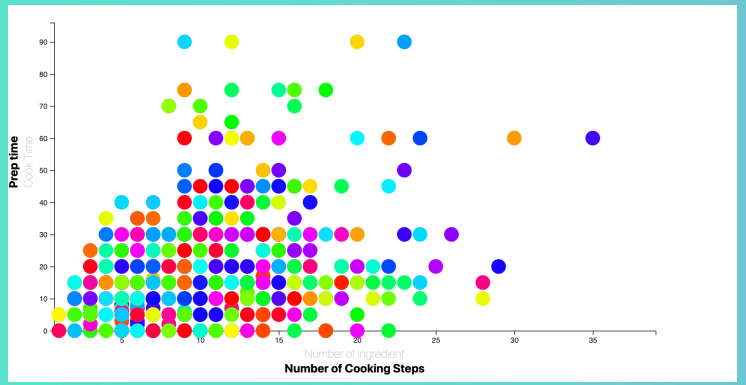
All 1588 data were
uploaded
and saved in MongoDB

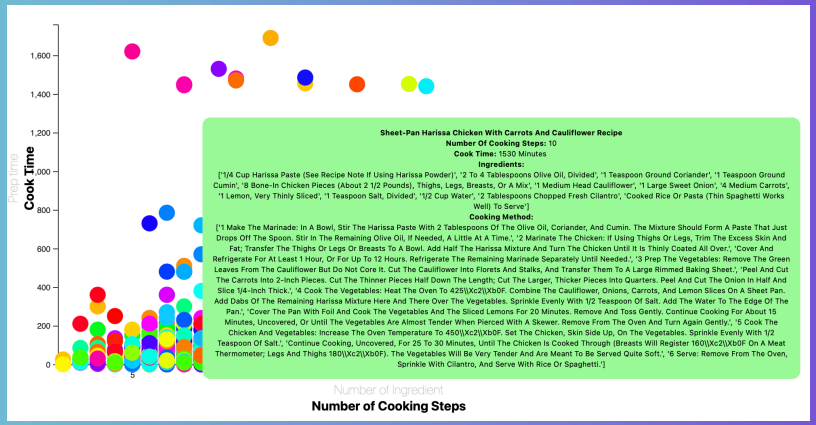


Total 1558 recipes were saved as csv and json format

HEESUNG'S CHART







Caitlin’s chart

A FEW WORDS FROM CAITLIN

CAITLIN'S CHART

FLASK 04

How do we present our findings

FLASK STRUCTURE

```python
import flask
from flask import Flask, jsonify, Response, render_template
from flask import redirect
import pymongo
from pymongo import MongoClient
from bson import ObjectId, json_util
import json

cluster = pymongo.MongoClient("mongodb+srv://group2:group2@cluster0.mpjcg.mongodb.net/<dbname>?retryWrites=true&w=majority")
db = cluster["simply_recipe"]
collection = db["recipes_collection"]

app = Flask(__name__)

# This route returns the team's index page
@app.route("/")
def home():
    return render_template('index.html')

# This route returns heesung's plot page of the team's website
@app.route("/heesung/")
def heesung():
    return redirect("https://heesung80.github.io/recipe/")

# This route returns caitlin's plot page of the team's website
@app.route("/caitlin")
def caitlin_plots():
    return render_template('inner-page_caitlin.html')

# This route returns all the recipe_collection data in JSON.
@app.route("/recipes", methods=["GET"])
def get_recipes():
    all_recipes = list(collection.find({}))
    return json.dumps(all_recipes, default=json_util.default)

if __name__ == "__main__":
    app.run(debug=True, host='0.0.0.0', port=port)
```

Dependencies

Connecting to our MongoDB Cluster

Rendering index page

Redirect – to Heesung's chart

Rendering an inner page

GET Request

Q&A

That's all folks!