WEBSCRAPPING NEW ZEALAND TRAVEL WEBSITE

I Putu Agastya Harta Pratama (472876)
Łukasz Brzoska (472892)

WHY THIS WEBSITE?

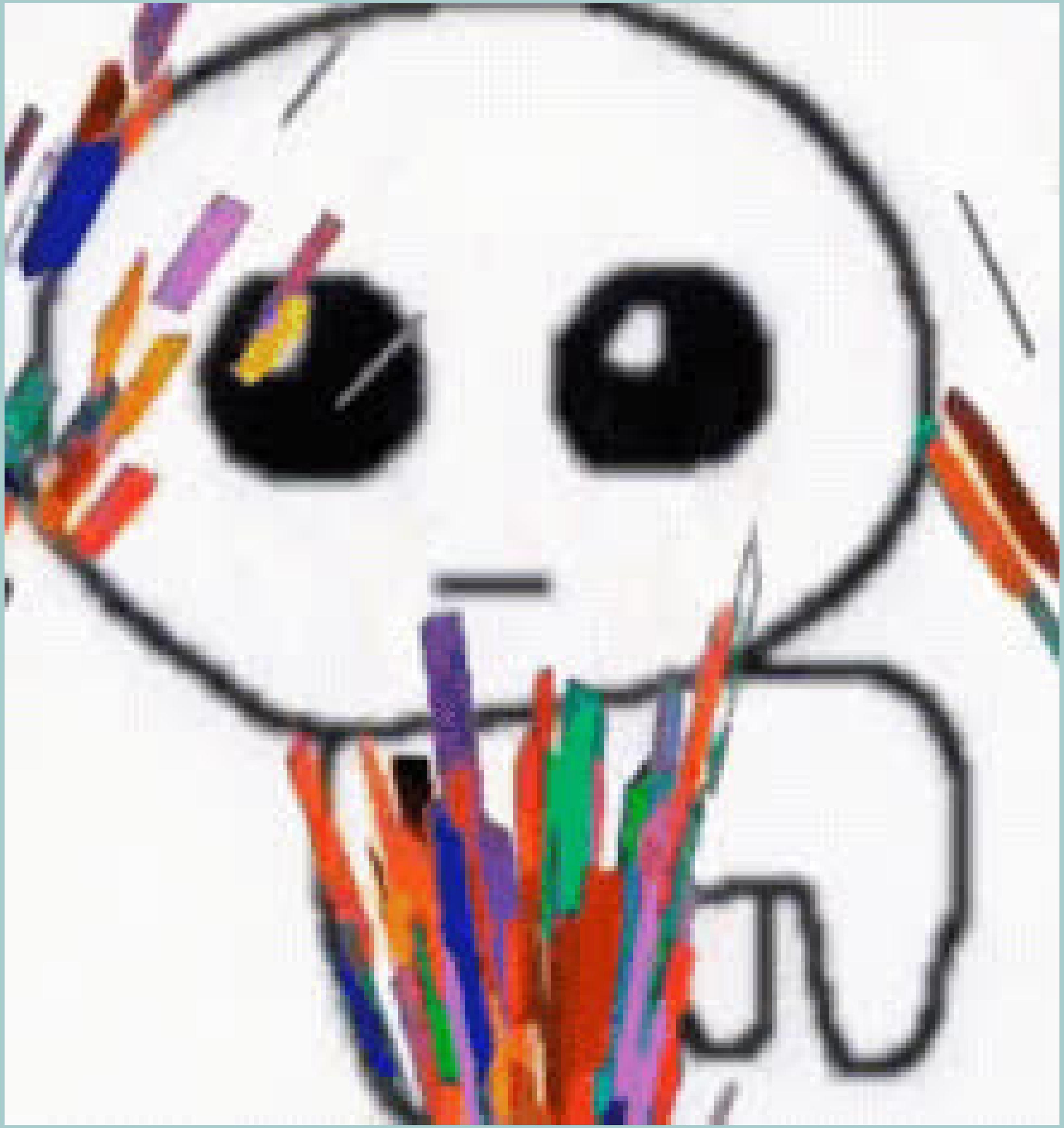
- Dynamic nature of this website allows us to fully showcase our webscrapping skills

- A wide variety of travel-related information about the country of New Zealand allowed us to scrape data on activieties in multiple cities that created an intresting dataset

- The scope of information contained in this website was really robust and we were also drawn by the design of the website

Were we allowed to scrape this website?

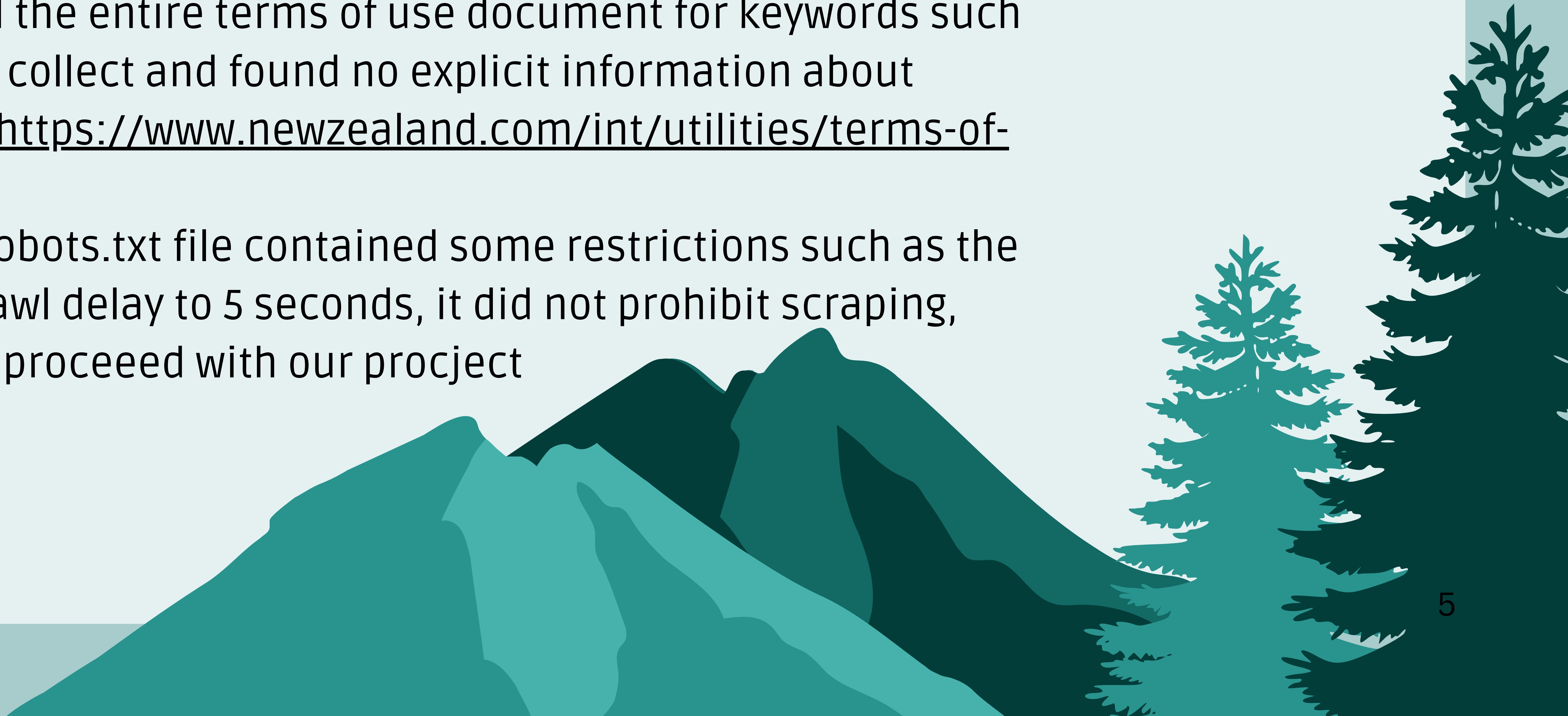
YES.



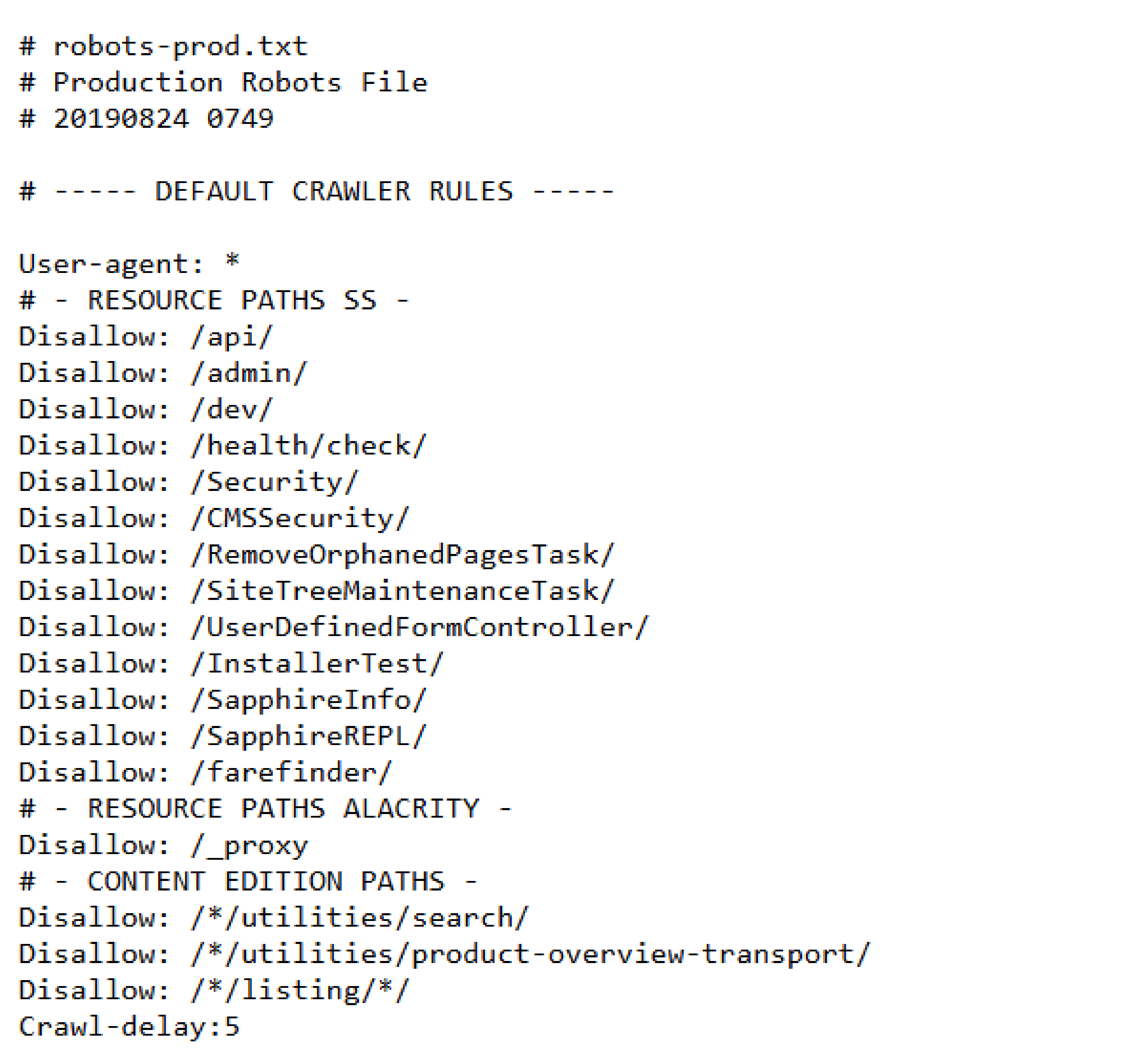
TERMS OF USE AND ROBOTS.TXT FILE

In order to check whether our project is within the legal and ethical norms, we checked bothe terms of use of the newzealand.com website as well as the robots.txt file:

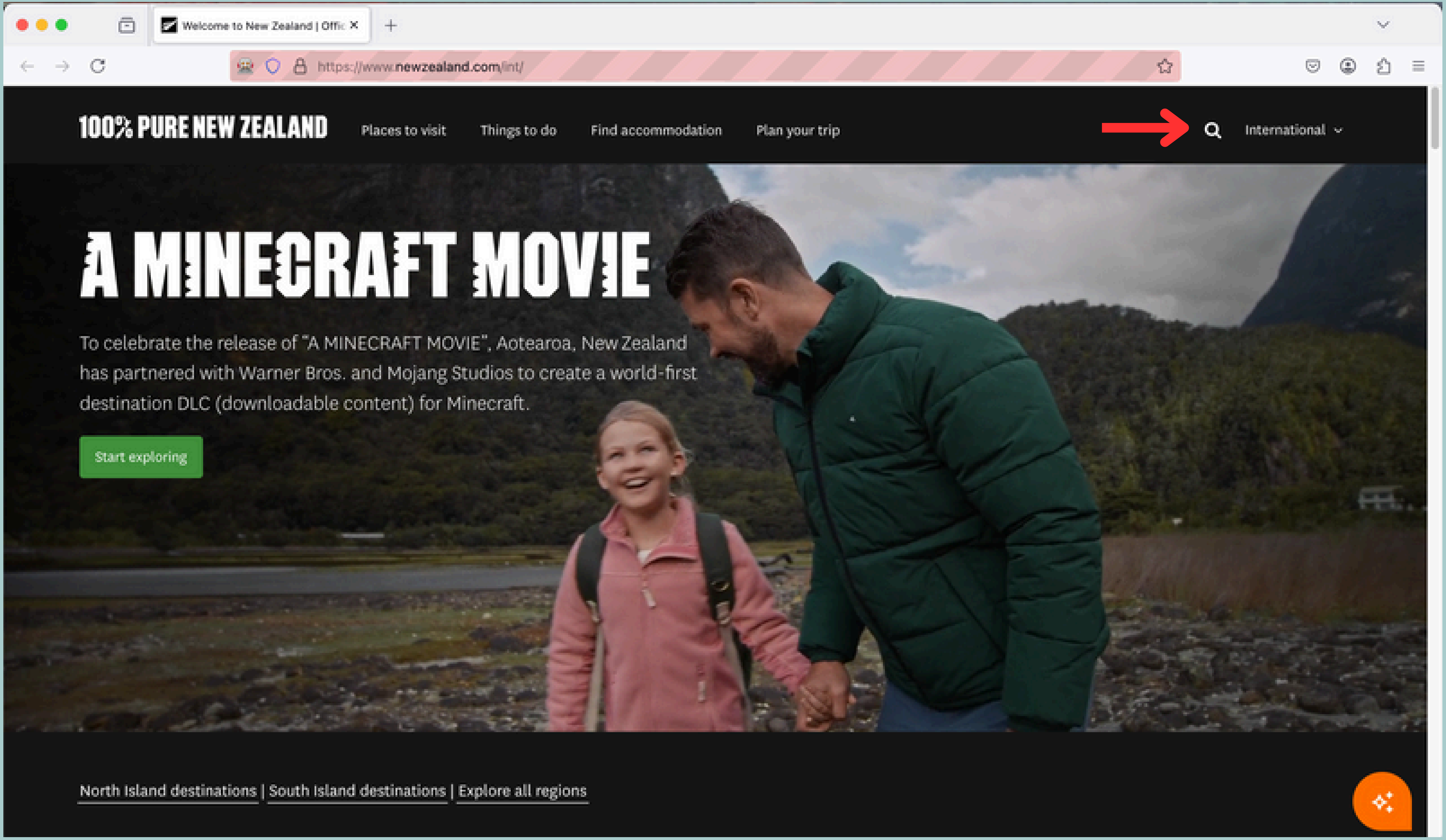
- Terms of use - We checked the entire terms of use document for keywords such as : robot, crawl, scrape or collect and found no explicit information about scraping being forbidden (https://www.newzealand.com/int/utilities/terms-of-use/)
- Robots.txt file - While the robots.txt file contained some restrictions such as the necessity of setting the crawl delay to 5 seconds, it did not prohibit scraping, therefore, we were able to proceeed with our procject



Contents of robots.txt



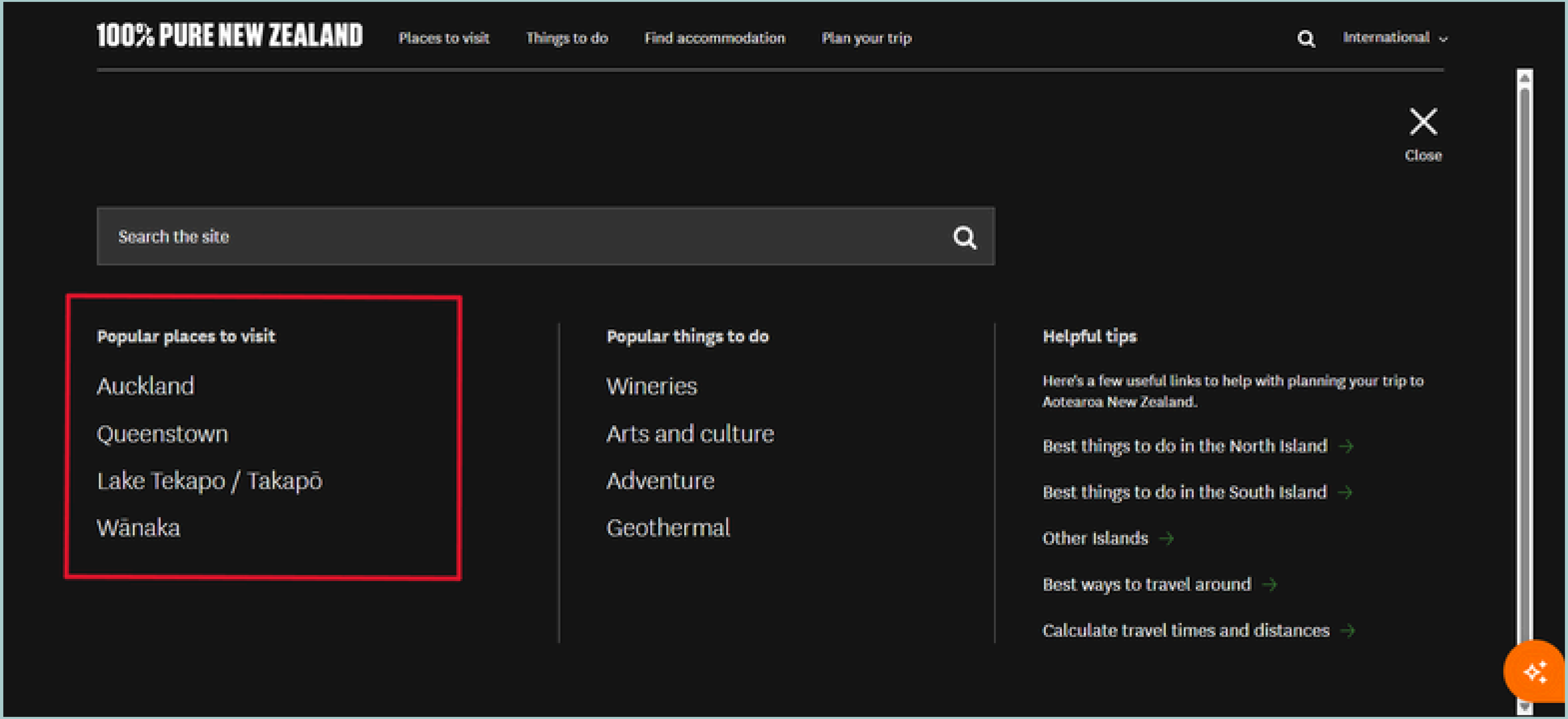
```
# robots-prod.txt
# Production Robots File
# 20190824 0749

# ----- DEFAULT CRAWLER RULES -----

User-agent: *
# - RESOURCE PATHS SS -
Disallow: /api/
Disallow: /admin/
Disallow: /dev/
Disallow: /health/check/
Disallow: /Security/
Disallow: /CMSSecurity/
Disallow: /RemoveOrphanedPagesTask/
Disallow: /SiteTreeMaintenanceTask/
Disallow: /UserDefinedFormController/
Disallow: /InstallerTest/
Disallow: /SapphireInfo/
Disallow: /SapphireREPL/
Disallow: /farefinder/
# - RESOURCE PATHS ALACRITY -
Disallow: /_proxy
# - CONTENT EDITION PATHS -
Disallow: /*/utilities/search/
Disallow: /*/utilities/product-overview-transport/
Disallow: /*/listing/*/
Crawl-delay:5
```


Welcome to New Zealand | Offic
https://www.newzealand.com/int/
100% PURE NEW ZEALAND
Places to visit
Things to do
Find accommodation
Plan your trip
International
Start exploring
North Island destinations | South Island destinations | Explore all regions

A MINECRAFT MOVIE

To celebrate the release of "A MINECRAFT MOVIE", Aotearoa, New Zealand has partnered with Warner Bros. and Mojang Studios to create a world-first destination DLC (downloadable content) for Minecraft.

WHAT WE SCRAPPED:



Popular places to visit

Auckland

Queenstown

Lake Tekapo / Takapō

Wānaka

Popular things to do

Wineries

Arts and culture

Adventure

Geothermal

Helpful tips

Here's a few useful links to help with planning your trip to Aotearoa New Zealand.

Best things to do in the North Island →

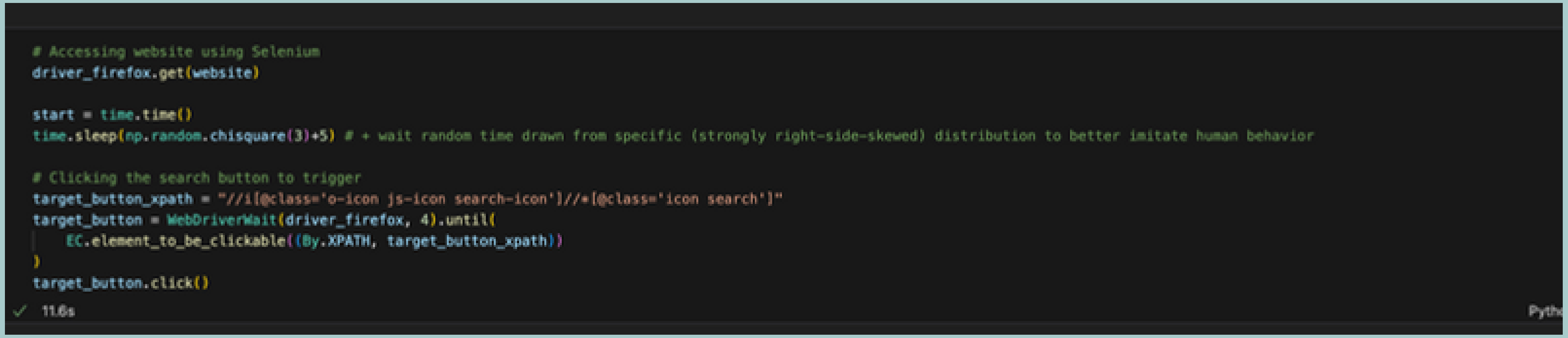
Best things to do in the South Island →

Other Islands →

Best ways to travel around →

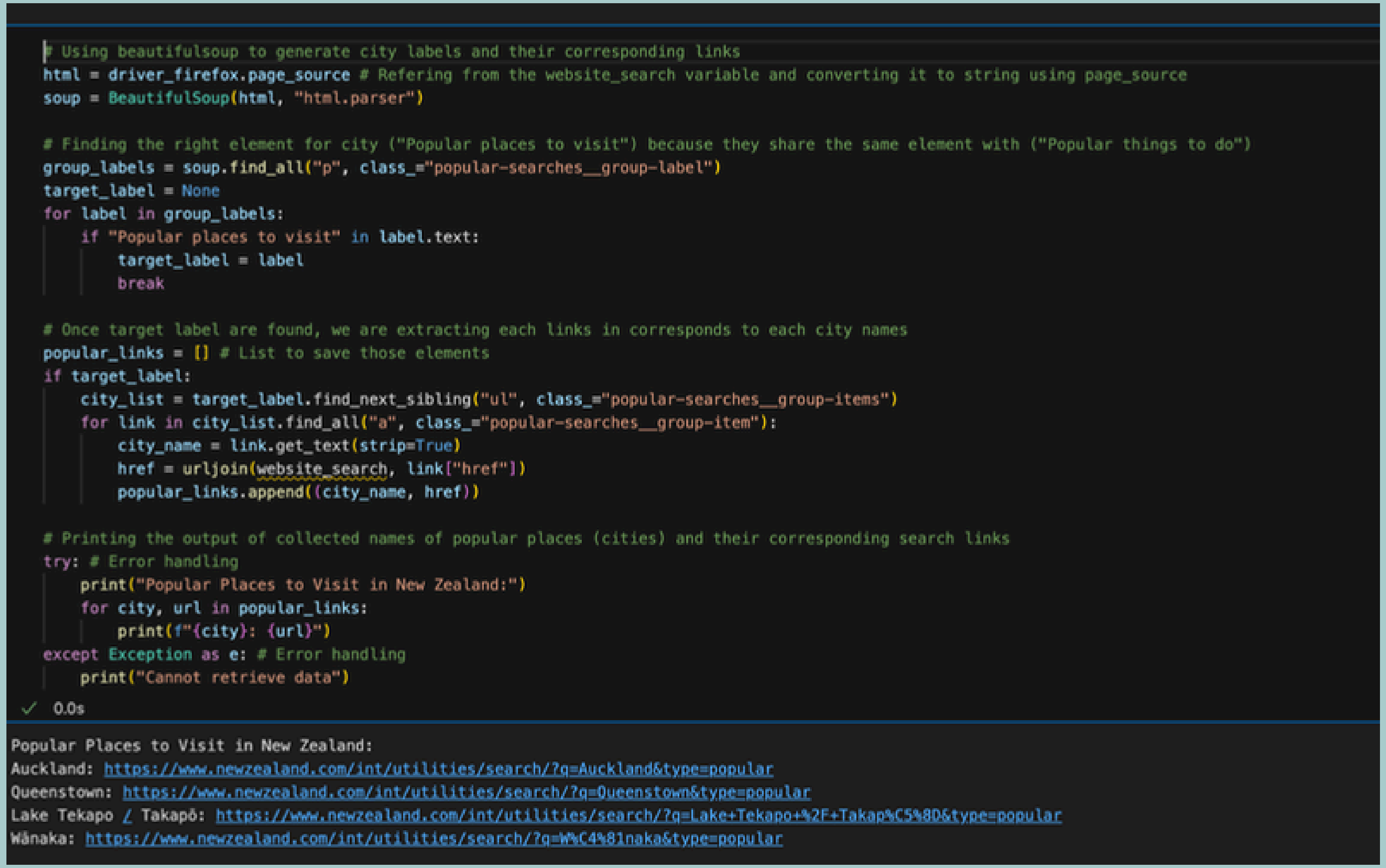
Calculate travel times and distances →

CODE SNIPPETS



```python
# Accessing website using Selenium
driver_firefox.get(website)

start = time.time()
time.sleep(np.random.chisquare(3)+5) # + wait random time drawn from specific (strongly right-side-skewed) distribution to better imitate human behavior

# Clicking the search button to trigger
target_button_xpath = "//i[@class='o-icon js-icon search-icon']//*[@class='icon search']"
target_button = WebDriverWait(driver_firefox, 4).until(
    EC.element_to_be_clickable((By.XPATH, target_button_xpath))
)
target_button.click()
```

✓ 11.6s

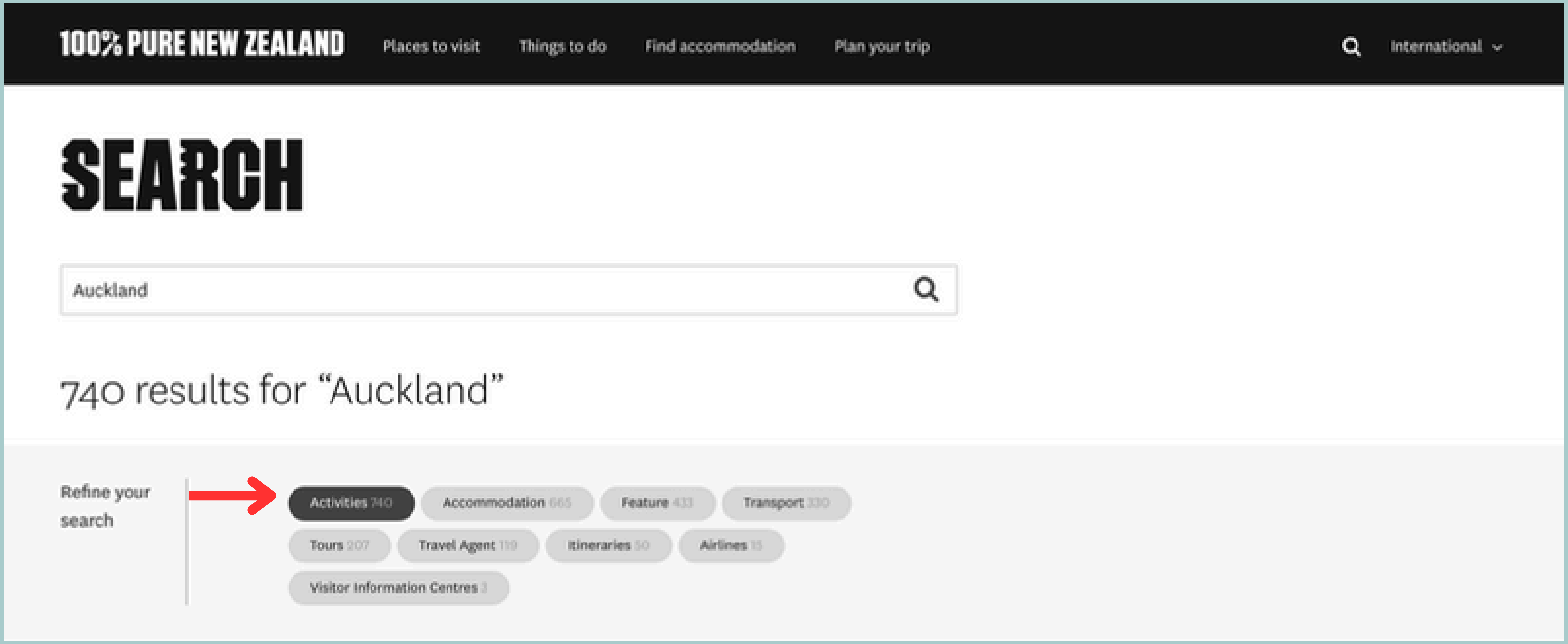


```python
# Using beautifulsoup to generate city labels and their corresponding links
html = driver_firefox.page_source # Refering from the website_search variable and converting it to string using page_source
soup = BeautifulSoup(html, "html.parser")

# Finding the right element for city ("Popular places to visit") because they share the same element with ("Popular things to do")
group_labels = soup.find_all("p", class_="popular-searches__group-label")
target_label = None
for label in group_labels:
    if "Popular places to visit" in label.text:
        target_label = label
        break

# Once target label are found, we are extracting each links in corresponds to each city names
popular_links = [] # List to save those elements
if target_label:
    city_list = target_label.find_next_sibling("ul", class_="popular-searches__group-items")
    for link in city_list.find_all("a", class_="popular-searches__group-item"):
        city_name = link.get_text(strip=True)
        href = urljoin(website_search, link["href"])
        popular_links.append((city_name, href))

# Printing the output of collected names of popular places (cities) and their corresponding search links
try: # Error handling
    print("Popular Places to Visit in New Zealand:")
    for city, url in popular_links:
        print(f"{city}: {url}")
except Exception as e: # Error handling
    print("Cannot retrieve data")
```

✓ 0.0s

```
Popular Places to Visit in New Zealand:
Auckland: https://www.newzealand.com/int/utilities/search/?q=Auckland&type=popular
Queenstown: https://www.newzealand.com/int/utilities/search/?q=Queenstown&type=popular
Lake Tekapo / Takapō: https://www.newzealand.com/int/utilities/search/?q=Lake+Tekapo+%2F+Takap%C5%8D&type=popular
Wānaka: https://www.newzealand.com/int/utilities/search/?q=W%C4%81naka&type=popular
```

WHAT WE SCRAPPED:

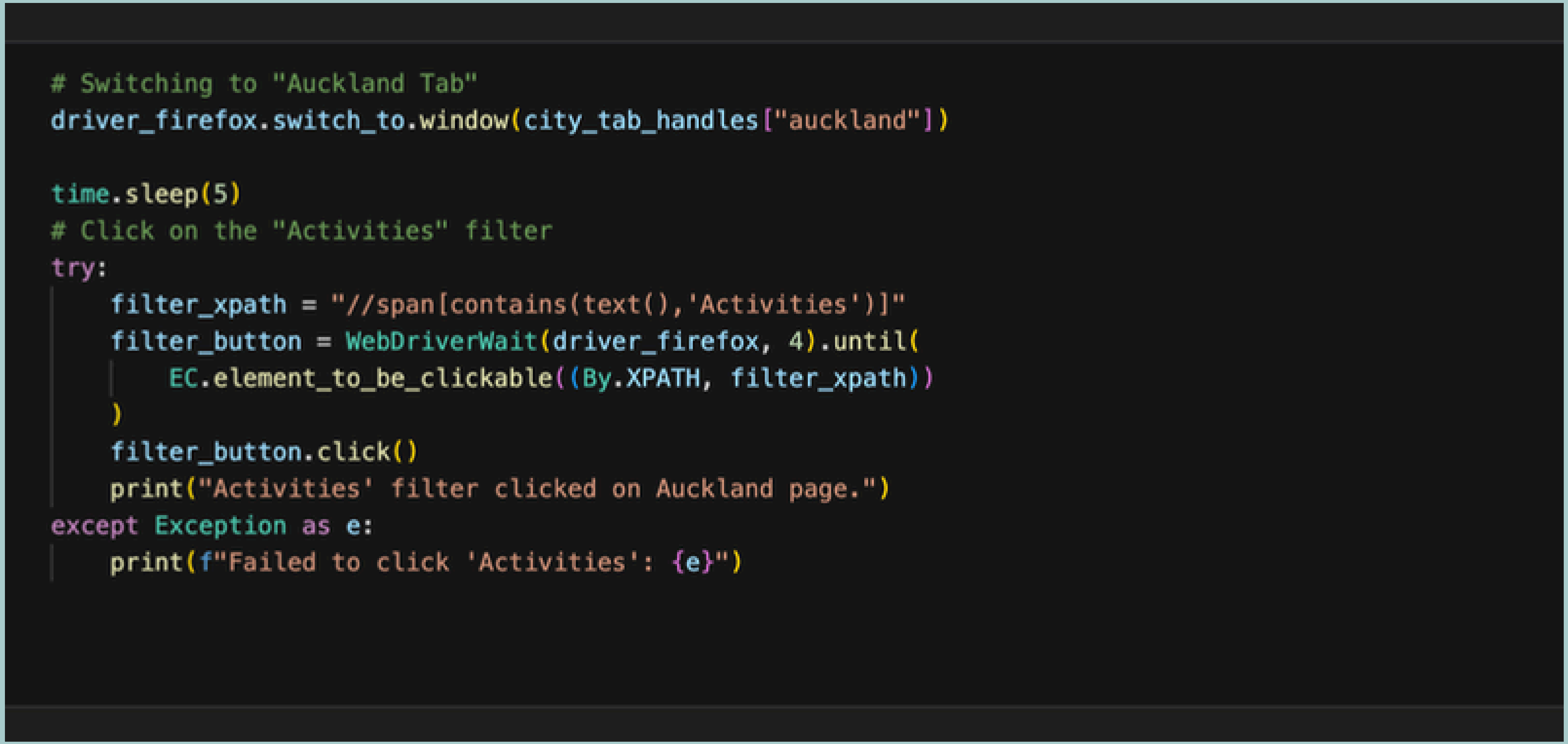


SEARCH

Auckland

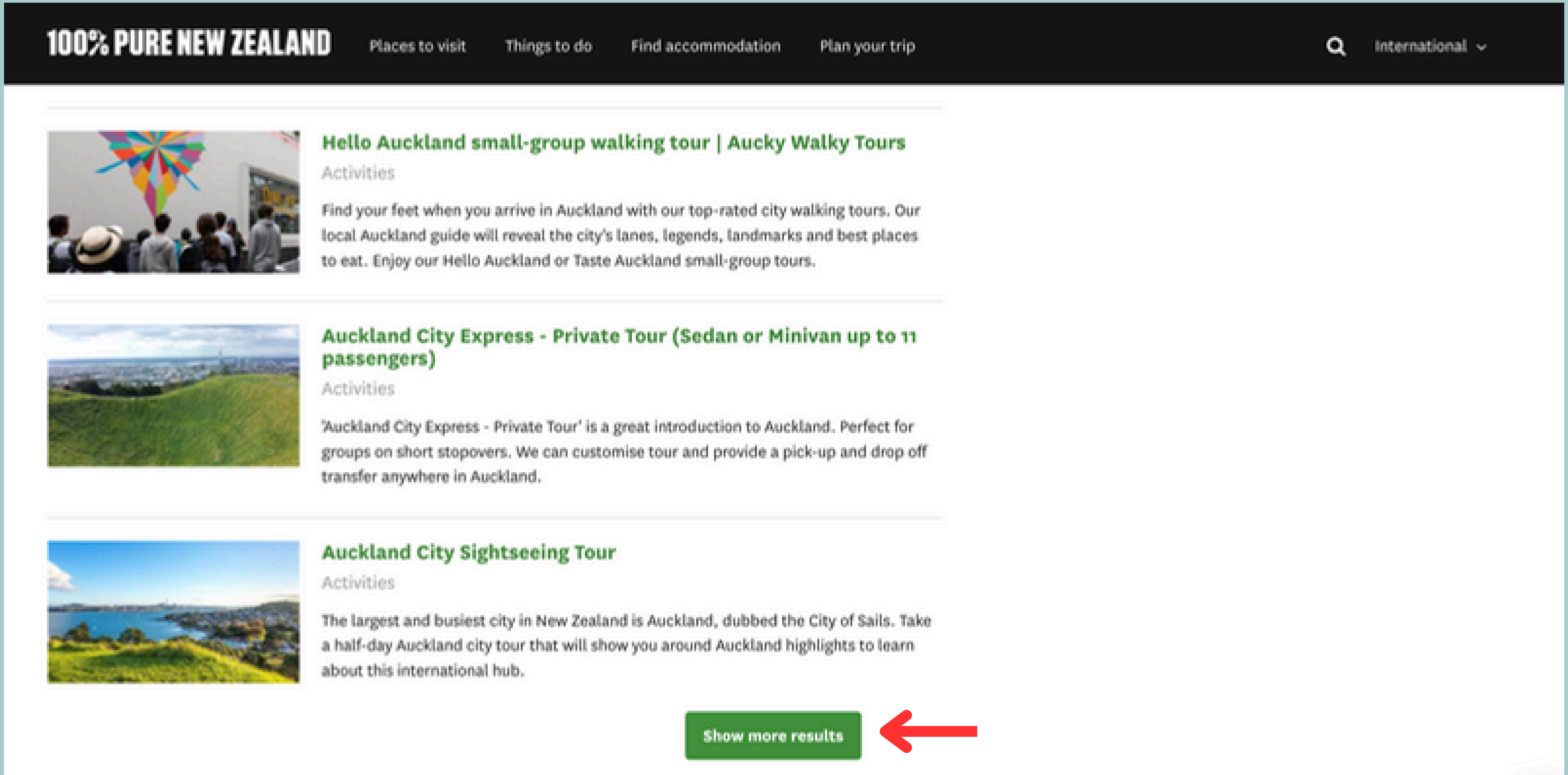
740 results for “Auckland”

WHAT WE SCRAPPED:



```python
# Switching to "Auckland Tab"
driver_firefox.switch_to.window(city_tab_handles["auckland"])

time.sleep(5)
# Click on the "Activities" filter
try:
    filter_xpath = "//span[contains(text(),'Activities')]"
    filter_button = WebDriverWait(driver_firefox, 4).until(
        EC.element_to_be_clickable((By.XPATH, filter_xpath))
    )
    filter_button.click()
    print("Activities' filter clicked on Auckland page.")
except Exception as e:
    print(f"Failed to click 'Activities': {e}")
```

WHAT WE SCRAPPED:



Hello Auckland small-group walking tour | Aucky Walky Tours

Activities

Find your feet when you arrive in Auckland with our top-rated city walking tours. Our local Auckland guide will reveal the city's lanes, legends, landmarks and best places to eat. Enjoy our Hello Auckland or Taste Auckland small-group tours.

Auckland City Express - Private Tour (Sedan or Minivan up to 11 passengers)

Activities

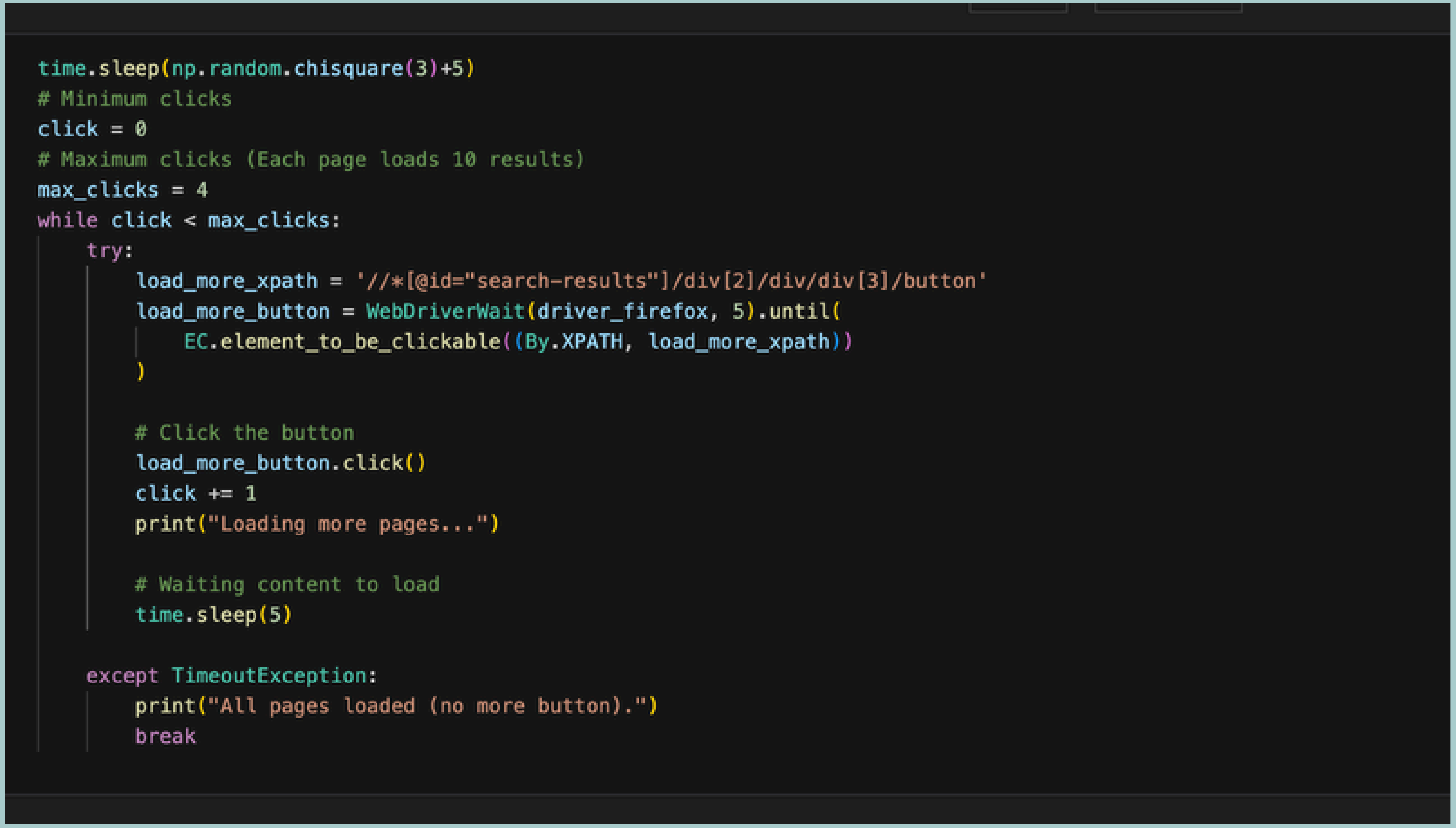
'Auckland City Express - Private Tour' is a great introduction to Auckland. Perfect for groups on short stopovers. We can customise tour and provide a pick-up and drop off transfer anywhere in Auckland.

Auckland City Sightseeing Tour

Activities

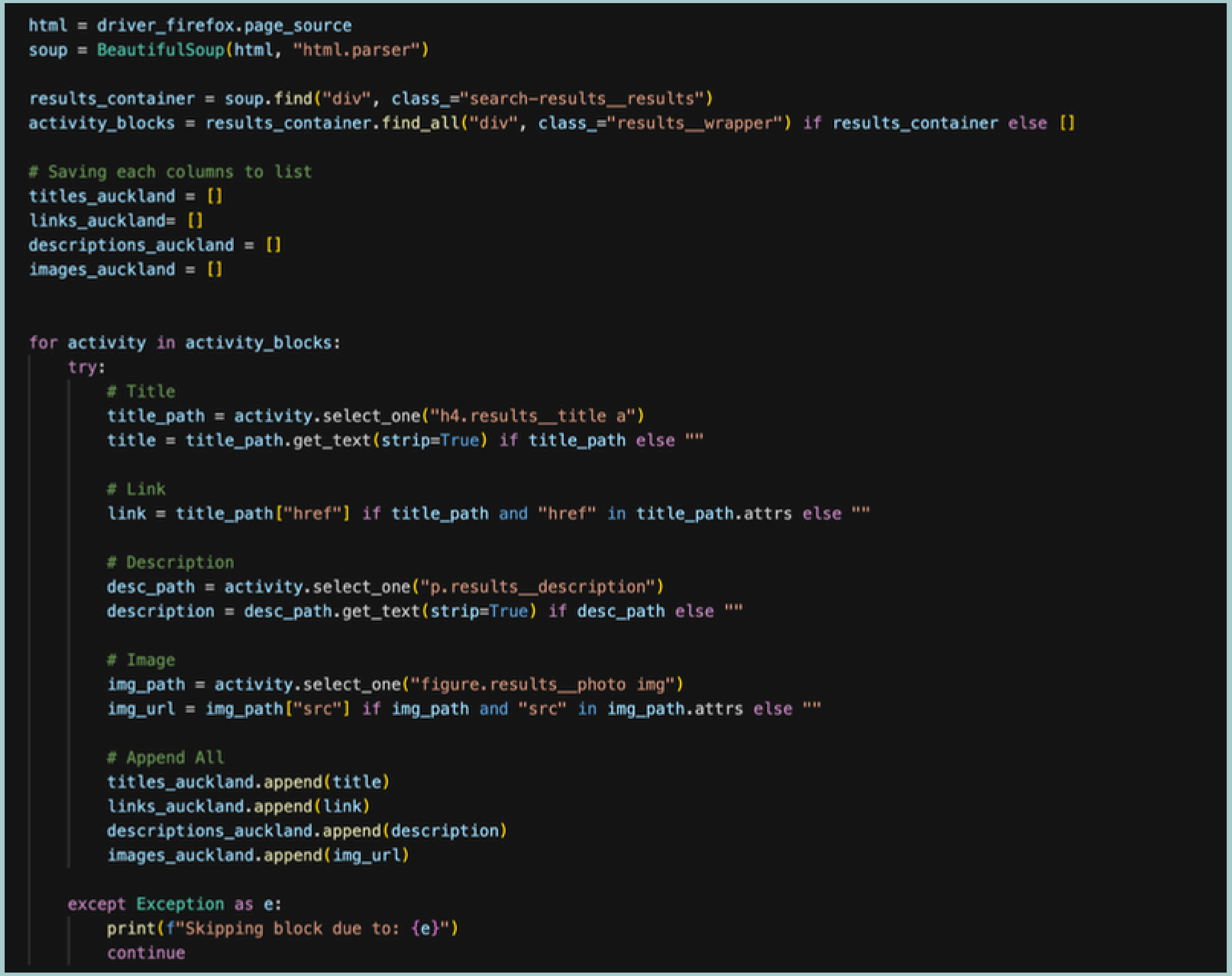
The largest and busiest city in New Zealand is Auckland, dubbed the City of Sails. Take a half-day Auckland city tour that will show you around Auckland highlights to learn about this international hub.

CODE SNIPPETS



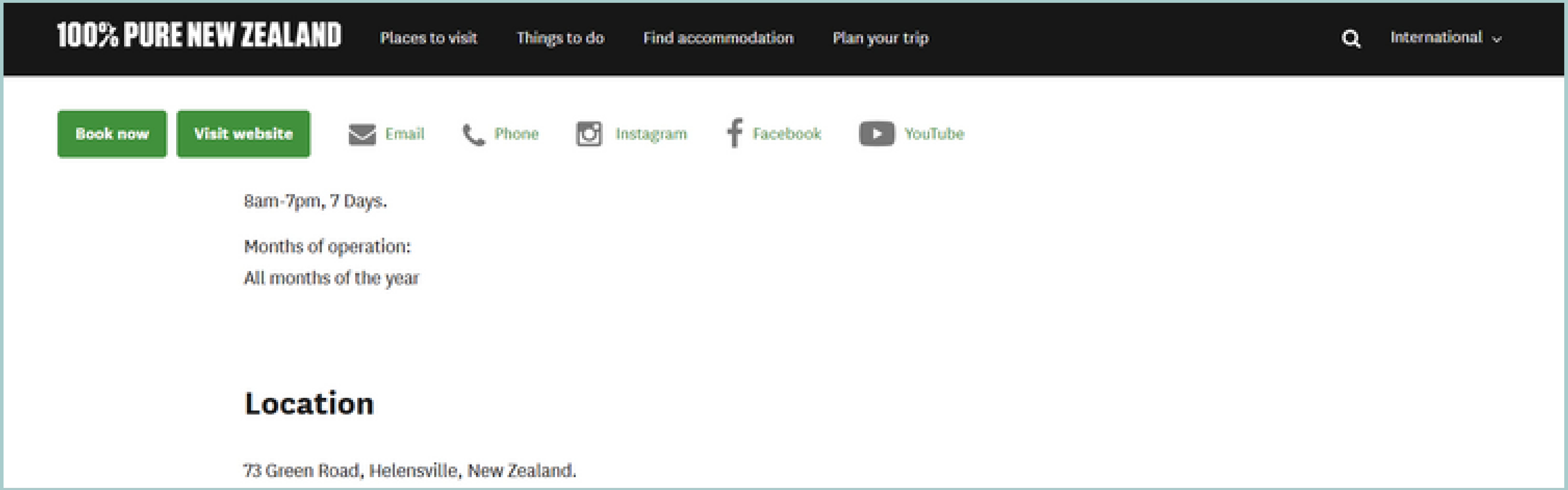
```
time.sleep(np.random.chisquare(3)+5)
# Minimum clicks
click = 0
# Maximum clicks (Each page loads 10 results)
max_clicks = 4
while click < max_clicks:
    try:
        load_more_xpath = '//*[@id="search-results"]/div[2]/div/div[3]/button'
        load_more_button = WebDriverWait(driver_firefox, 5).until(
            EC.element_to_be_clickable((By.XPATH, load_more_xpath))
        )

        # Click the button
        load_more_button.click()
        click += 1
        print("Loading more pages...")

        # Waiting content to load
        time.sleep(5)

    except TimeoutException:
        print("All pages loaded (no more button).")
        break
```

CODE SNIPPETS



```python
html = driver_firefox.page_source
soup = BeautifulSoup(html, "html.parser")

results_container = soup.find("div", class_="search-results__results")
activity_blocks = results_container.find_all("div", class_="results__wrapper") if results_container else []

# Saving each columns to list
titles_auckland = []
links_auckland= []
descriptions_auckland = []
images_auckland = []


for activity in activity_blocks:
    try:
        # Title
        title_path = activity.select_one("h4.results__title a")
        title = title_path.get_text(strip=True) if title_path else ""

        # Link
        link = title_path["href"] if title_path and "href" in title_path.attrs else ""

        # Description
        desc_path = activity.select_one("p.results__description")
        description = desc_path.get_text(strip=True) if desc_path else ""

        # Image
        img_path = activity.select_one("figure.results__photo img")
        img_url = img_path["src"] if img_path and "src" in img_path.attrs else ""

        # Append All
        titles_auckland.append(title)
        links_auckland.append(link)
        descriptions_auckland.append(description)
        images_auckland.append(img_url)

    except Exception as e:
        print(f"Skipping block due to: {e}")
        continue
```

WHAT WE SCRAPPED:



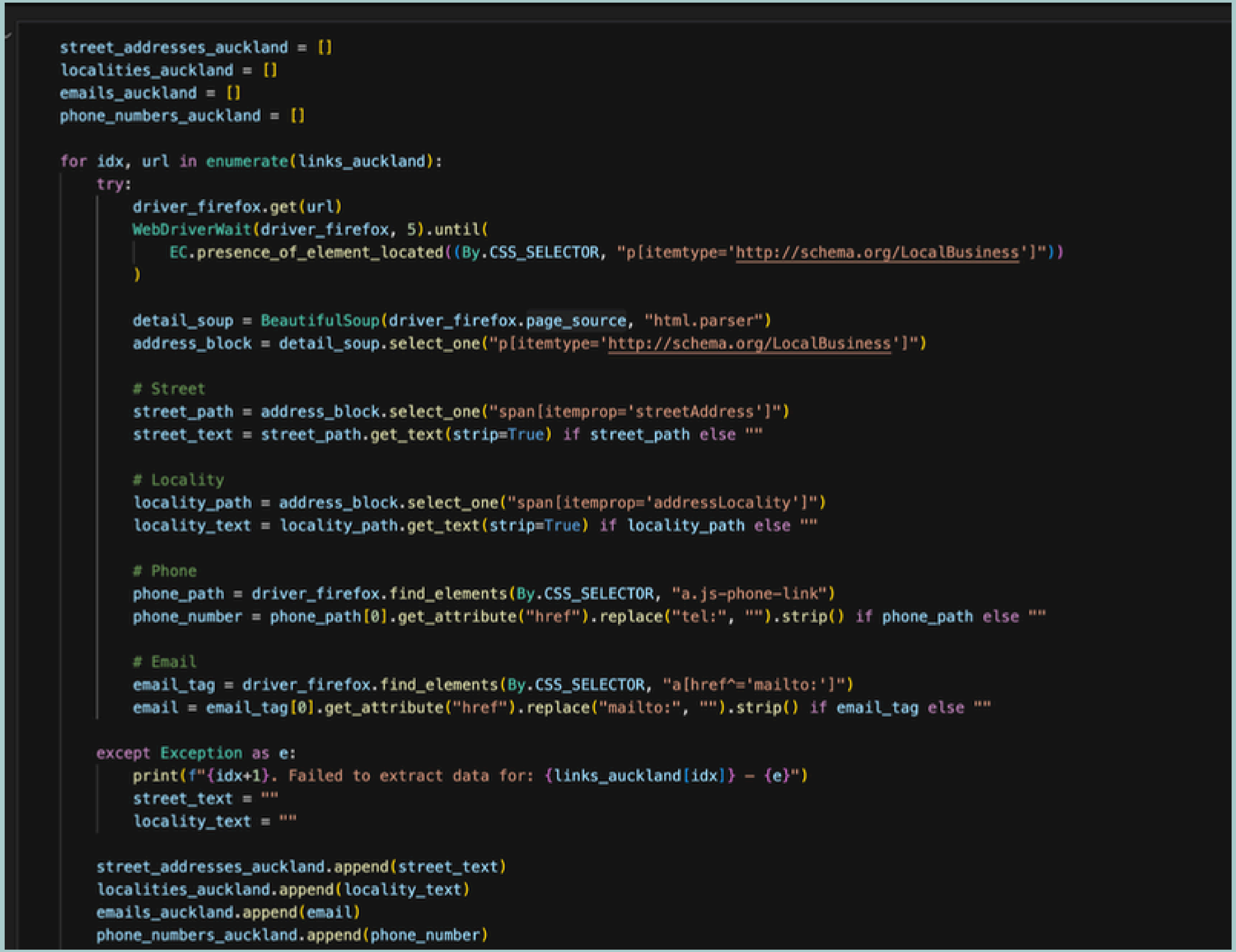
8am-7pm, 7 Days.

Months of operation:
All months of the year

Location

73 Green Road, Helensville, New Zealand.

CODE SNIPPETS

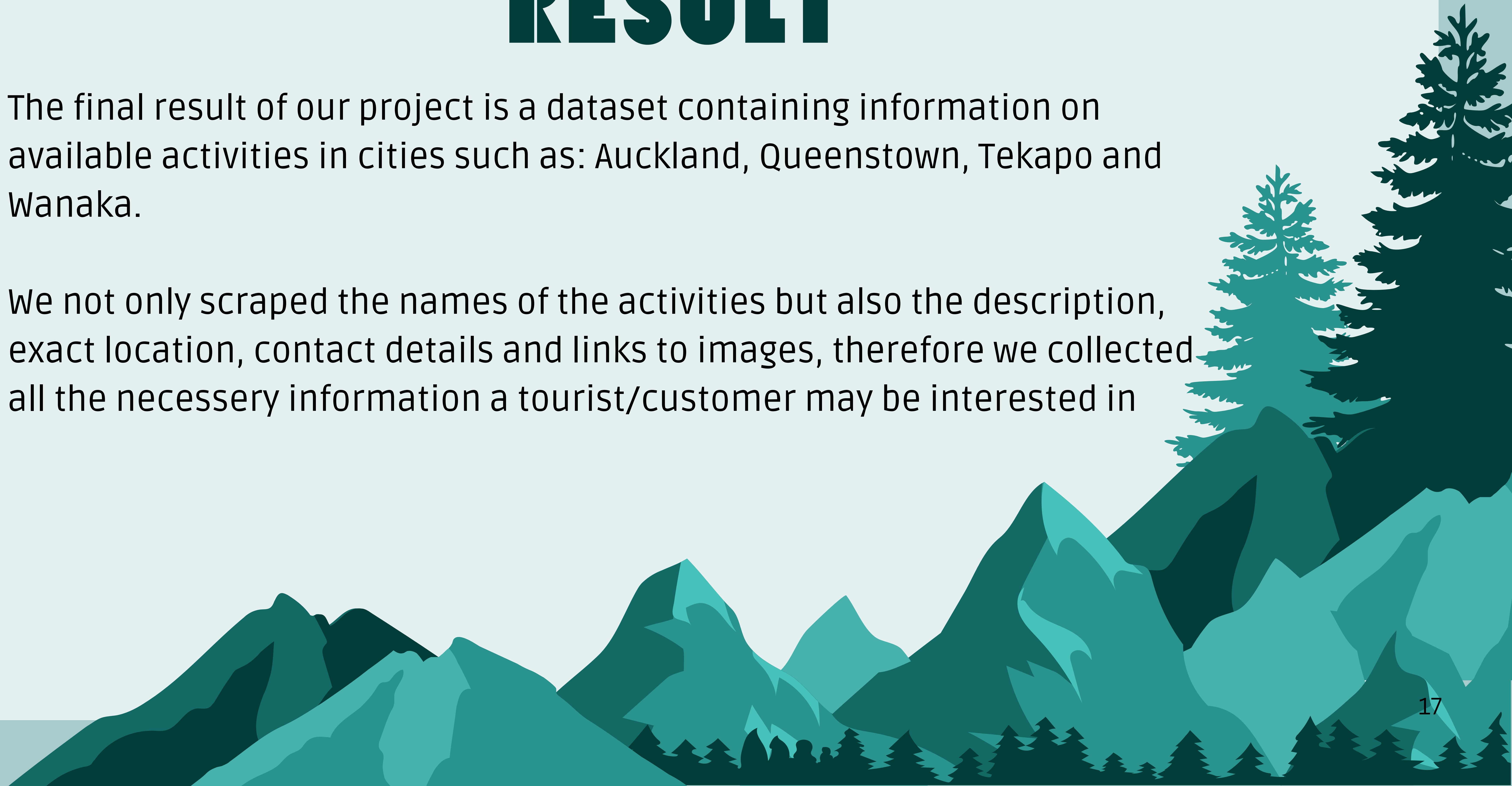


```python
street_addresses_auckland = []
localities_auckland = []
emails_auckland = []
phone_numbers_auckland = []

for idx, url in enumerate(links_auckland):
    try:
        driver_firefox.get(url)
        WebDriverWait(driver_firefox, 5).until(
            EC.presence_of_element_located((By.CSS_SELECTOR, "p[itemtype='http://schema.org/LocalBusiness']"))
        )

        detail_soup = BeautifulSoup(driver_firefox.page_source, "html.parser")
        address_block = detail_soup.select_one("p[itemtype='http://schema.org/LocalBusiness']")

        # Street
        street_path = address_block.select_one("span[itemprop='streetAddress']")
        street_text = street_path.get_text(strip=True) if street_path else ""

        # Locality
        locality_path = address_block.select_one("span[itemprop='addressLocality']")
        locality_text = locality_path.get_text(strip=True) if locality_path else ""

        # Phone
        phone_path = driver_firefox.find_elements(By.CSS_SELECTOR, "a.js-phone-link")
        phone_number = phone_path[0].get_attribute("href").replace("tel:", "").strip() if phone_path else ""

        # Email
        email_tag = driver_firefox.find_elements(By.CSS_SELECTOR, "a[href^='mailto:']")
        email = email_tag[0].get_attribute("href").replace("mailto:", "").strip() if email_tag else ""

    except Exception as e:
        print(f"{idx+1}. Failed to extract data for: {links_auckland[idx]} - {e}")
        street_text = ""
        locality_text = ""

    street_addresses_auckland.append(street_text)
    localities_auckland.append(locality_text)
    emails_auckland.append(email)
    phone_numbers_auckland.append(phone_number)
```

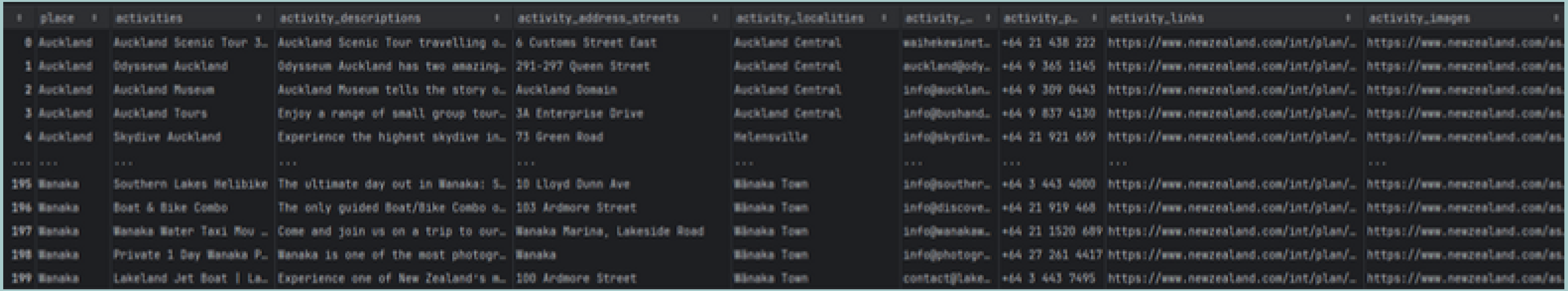
RESULT

The final result of our project is a dataset containing information on available activities in cities such as: Auckland, Queenstown, Tekapo and Wanaka.

We not only scraped the names of the activities but also the description, exact location, contact details and links to images, therefore we collected all the necessery information a tourist/customer may be interested in



SAMPLE OUTPUT:



	place	activities	activity_descriptions	activity_address_streets	activity_localities	activity_…	activity_p…	activity_links	activity_images
0	Auckland	Auckland Scenic Tour 3…	Auckland Scenic Tour travelling o…	6 Customs Street East	Auckland Central	waihekewinet…	+64 21 438 222	https://www.newzealand.com/int/plan/…	https://www.newzealand.com/as…
1	Auckland	Odysseum Auckland	Odysseum Auckland has two amazing…	291-297 Queen Street	Auckland Central	auckland@ody…	+64 9 365 1145	https://www.newzealand.com/int/plan/…	https://www.newzealand.com/as…
2	Auckland	Auckland Museum	Auckland Museum tells the story o…	Auckland Domain	Auckland Central	info@aucklan…	+64 9 309 0443	https://www.newzealand.com/int/plan/…	https://www.newzealand.com/as…
3	Auckland	Auckland Tours	Enjoy a range of small group tour…	3A Enterprise Drive	Auckland Central	info@bushand…	+64 9 837 4130	https://www.newzealand.com/int/plan/…	https://www.newzealand.com/as…
4	Auckland	Skydive Auckland	Experience the highest skydive in…	73 Green Road	Helensville	info@skydive…	+64 21 921 659	https://www.newzealand.com/int/plan/…	https://www.newzealand.com/as…
...	...	...	...	...	...	...	...	...	...
195	Wanaka	Southern Lakes Helibike	The ultimate day out in Wanaka: S…	10 Lloyd Dunn Ave	Wānaka Town	info@souther…	+64 3 443 4000	https://www.newzealand.com/int/plan/…	https://www.newzealand.com/as…
196	Wanaka	Boat & Bike Combo	The only guided Boat/Bike Combo o…	103 Ardmore Street	Wānaka Town	info@discove…	+64 21 919 468	https://www.newzealand.com/int/plan/…	https://www.newzealand.com/as…
197	Wanaka	Wanaka Water Taxi Mou …	Come and join us on a trip to our…	Wanaka Marina, Lakeside Road	Wānaka Town	info@wanakaw…	+64 21 1520 689	https://www.newzealand.com/int/plan/…	https://www.newzealand.com/as…
198	Wanaka	Private 1 Day Wanaka P…	Wanaka is one of the most photogr…	Wanaka	Wānaka Town	info@photogr…	+64 27 261 4417	https://www.newzealand.com/int/plan/…	https://www.newzealand.com/as…
199	Wanaka	Lakeland Jet Boat | La…	Experience one of New Zealand's m…	100 Ardmore Street	Wānaka Town	contact@lake…	+64 3 443 7495	https://www.newzealand.com/int/plan/…	https://www.newzealand.com/as…

CONCLUSION

- Selecting a dynamic and advanced website that can be legally and ethically scrapped
- Implementing both Beautiful Soup and Selenium in order to navigate the website
- Scraping the data and building a dataset from retrieved data

THANK YOU FOR YOUR ATTENTION