Web Scraping Project

Example - Kijiji

Cats & Kittens for Rehoming

Credits: Balifilm/ Shutterstock.com

June 2022

Photo credits: Kijiji and Pusha The Cat

Agenda

- Motivation
- Data: Web Scraping and Cleaning
- Data Analysis
- Conclusions
- Challenges
- Next Steps

Photo credits: Kijiji

Motivation

I have a cat

Are you looking at other cats on the Internet?
How dare you?

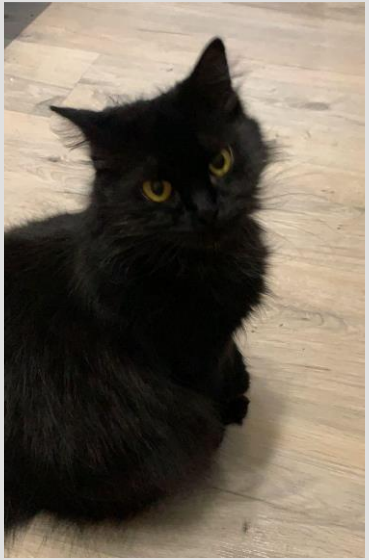
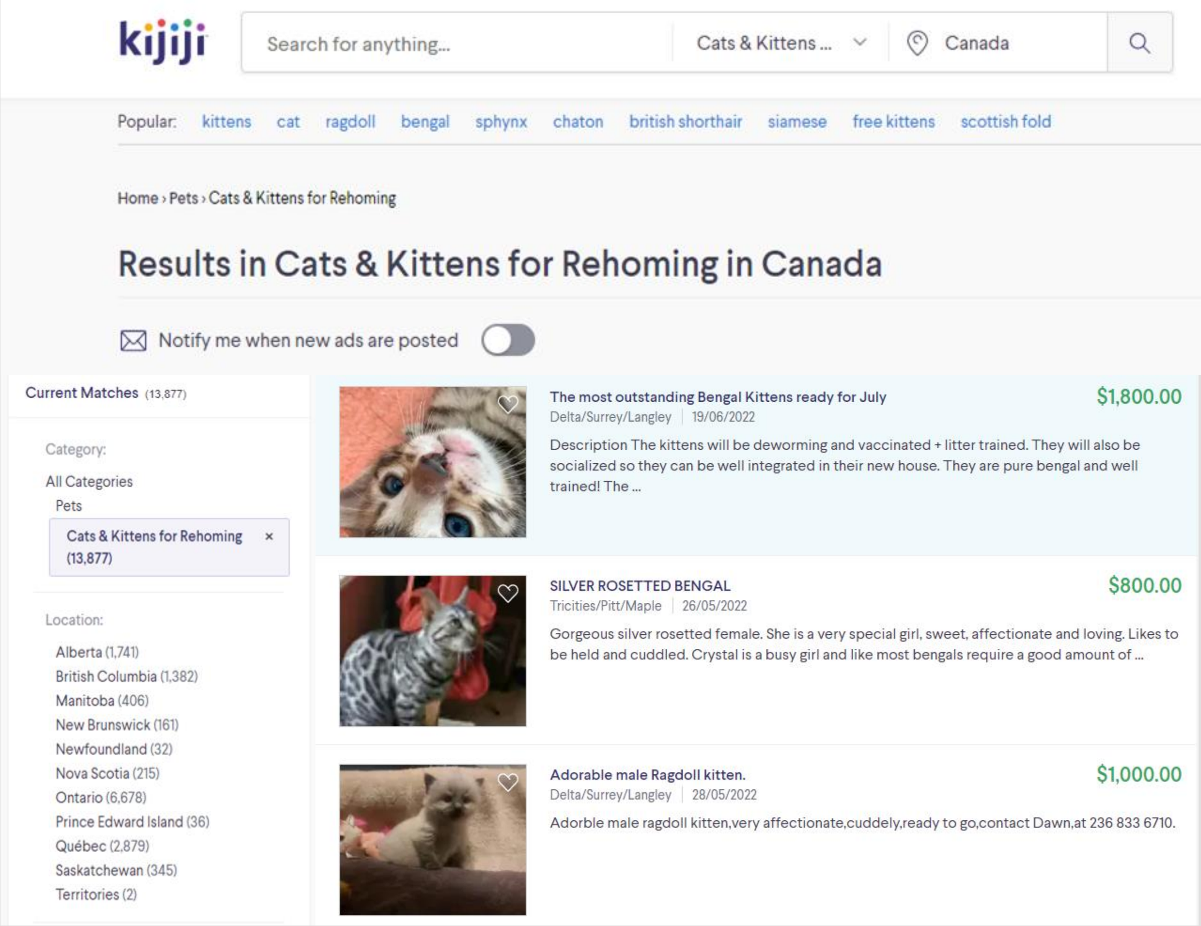


Photo credits: Pusha The Cat

Website



Popular: kittens cat ragdoll bengal sphynx chaton british shorthair siamese free kittens scottish fold

Home › Pets › Cats & Kittens for Rehoming

Post an Ad in this category

Results in Cats & Kittens for Rehoming in Canada

Showing 1 - 40 of 249 results

Notify me when new ads are posted

Posted: newest first

Current Matches (13,877)

Category:

All Categories

Pets

Cats & Kittens for Rehoming (13,877) ×

Location:

Alberta	(1,741)
British Columbia	(1,382)
Manitoba	(406)
New Brunswick	(161)
Newfoundland	(32)
Nova Scotia	(215)
Ontario	(6,678)
Prince Edward Island	(36)
Québec	(2,879)
Saskatchewan	(345)
Territories	(2)

The most outstanding Bengal Kittens ready for July $1,800.00

Delta/Surrey/Langley | 19/06/2022

Description The kittens will be deworming and vaccinated + litter trained. They will also be socialized so they can be well integrated in their new house. They are pure bengal and well trained! The ...

SILVER ROSETTED BENGAL $800.00

Tricities/Pitt/Maple | 26/05/2022

Gorgeous silver rosetted female. She is a very special girl, sweet, affectionate and loving. Likes to be held and cuddled. Crystal is a busy girl and like most bengals require a good amount of ...

Adorable male Ragdoll kitten. $1,000.00

Delta/Surrey/Langley | 28/05/2022

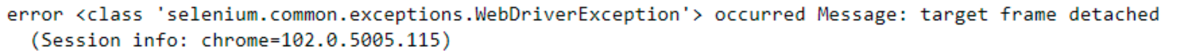
Adorble male ragdoll kitten,very affectionate,cuddely,ready to go,contact Dawn,at 236 833 6710.

Scraping Process

- Selenium was used to download a separate page for every listing. Then, BeautifulSoup scraped the data from the page;

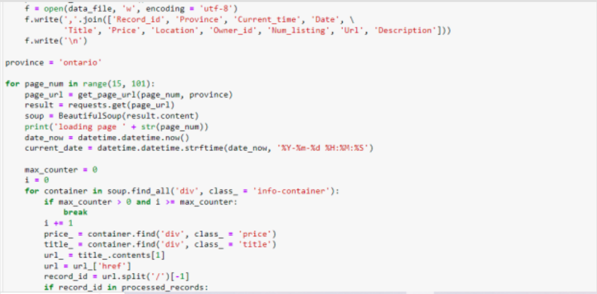
- Scraping process was going province by province;

- Approximately, 5% of ads failed to scrap (Error message is below)



```
error <class 'selenium.common.exceptions.WebDriverException'> occurred Message: target frame detached
  (Session info: chrome=102.0.5005.115)
```

Scraping Code Fragments

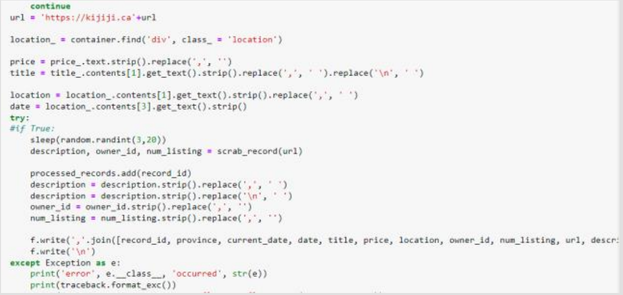


```
    f = open(data_file, 'w', encoding = 'utf-8')
    f.write(','.join(['Record_id', 'Province', 'Current_time', 'Date', \
            'Title', 'Price', 'Location', 'Owner_id', 'Num_listing', 'Url', 'Description']))
    f.write('\n')

province = 'ontario'

for page_num in range(15, 101):
    page_url = get_page_url(page_num, province)
    result = requests.get(page_url)
    soup = BeautifulSoup(result.content)
    print('loading page ' + str(page_num))
    date_now = datetime.datetime.now()
    current_date = datetime.datetime.strftime(date_now, '%Y-%m-%d %H:%M:%S')

    max_counter = 0
    i = 0
    for container in soup.find_all('div', class_ = 'info-container'):
        if max_counter > 0 and i >= max_counter:
            break
        i += 1
        price_ = container.find('div', class_ = 'price')
        title_ = container.find('div', class_ = 'title')
        url_ = title_.contents[1]
        url = url_['href']
        record_id = url.split('/')[-1]
        if record_id in processed_records:
```



```
    continue
url = 'https://kijiji.ca'+url

location_ = container.find('div', class_ = 'location')

price = price_.text.strip().replace(',', '')
title = title_.contents[1].get_text().strip().replace(',', ' ').replace('\n', ' ')

location = location_.contents[1].get_text().strip().replace(',', ' ')
date = location_.contents[3].get_text().strip()
try:
#if True:
    sleep(random.randint(3,20))
    description, owner_id, num_listing = scrab_record(url)

    processed_records.add(record_id)
    description = description.strip().replace(',', ' ')
    description = description.strip().replace('\n', ' ')
    owner_id = owner_id.strip().replace(',', '')
    num_listing = num_listing.strip().replace(',', '')

    f.write(','.join([record_id, province, current_date, date, title, price, location, owner_id, num_listing, url, descr
    f.write('\n')
except Exception as e:
    print('error', e.__class__, 'occurred', str(e))
    print(traceback.format_exc())
```

Photo credits: Kijiji

Data Cleaning

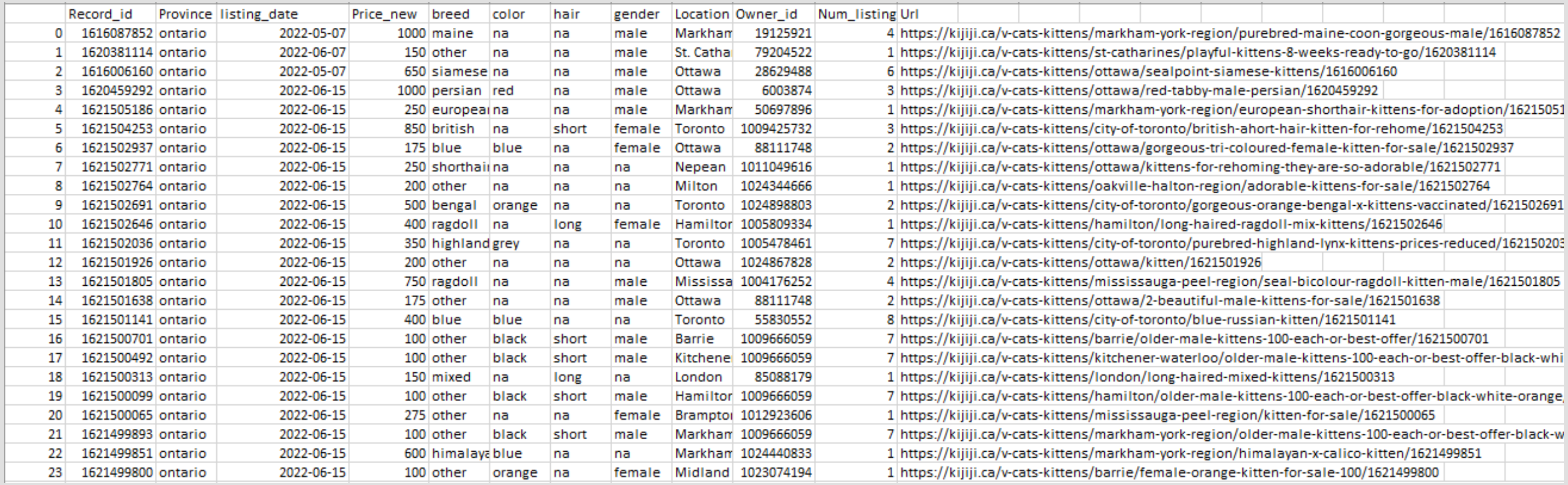
- Special code for data cleaning was created

- Breed, color, hair and gender data were extracted from both ad’s title and description

- Total data record: 3,354 rows

- Listings from Quebec were excluded from analysis

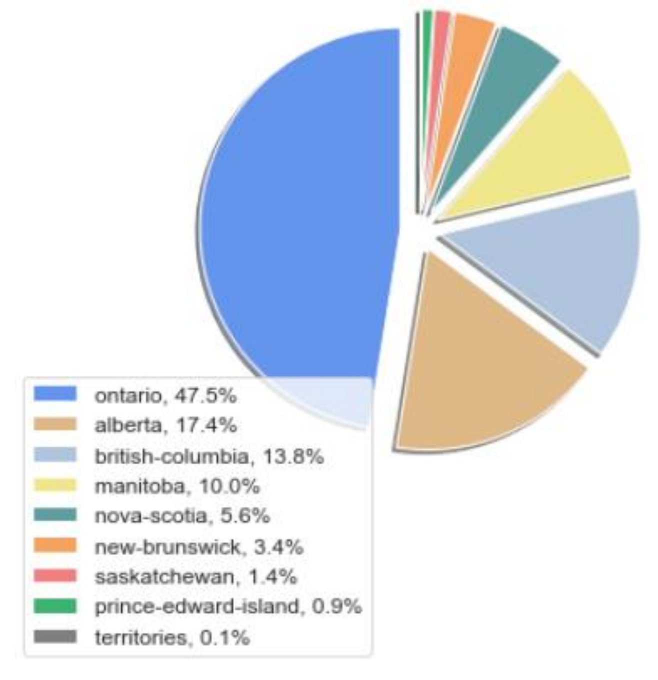
Clean Data Sample

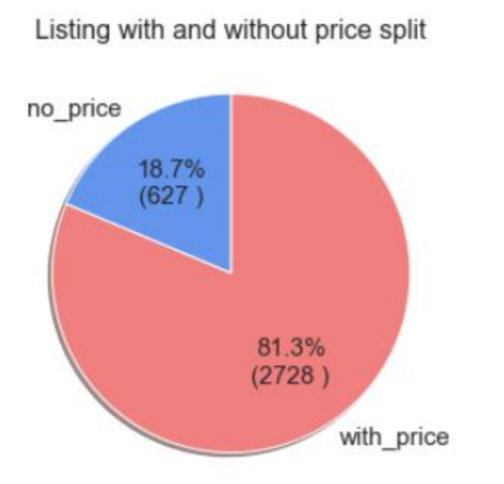


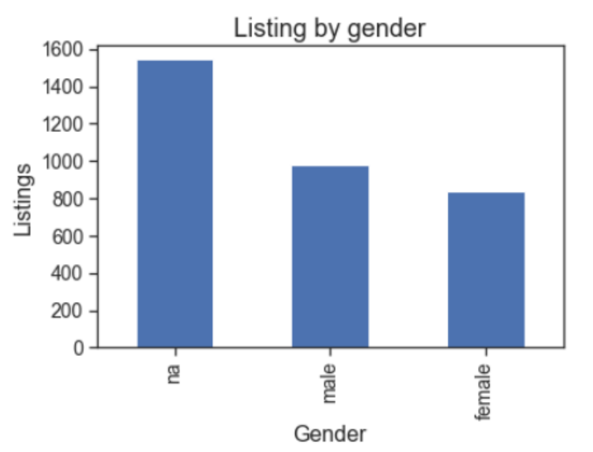
	Record_id	Province	listing_date	Price_new	breed	color	hair	gender	Location	Owner_id	Num_listing	Url
0	1616087852	ontario	2022-05-07	1000	maine	na	na	male	Markham	19125921	4	https://kijiji.ca/v-cats-kittens/markham-york-region/purebred-maine-coon-gorgeous-male/1616087852
1	1620381114	ontario	2022-06-07	150	other	na	na	male	St. Catha	79204522	1	https://kijiji.ca/v-cats-kittens/st-catharines/playful-kittens-8-weeks-ready-to-go/1620381114
2	1616006160	ontario	2022-05-07	650	siamese	na	na	male	Ottawa	28629488	6	https://kijiji.ca/v-cats-kittens/ottawa/sealpoint-siamese-kittens/1616006160
3	1620459292	ontario	2022-06-15	1000	persian	red	na	male	Ottawa	6003874	3	https://kijiji.ca/v-cats-kittens/ottawa/red-tabby-male-persian/1620459292
4	1621505186	ontario	2022-06-15	250	europea	na	na	male	Markham	50697896	1	https://kijiji.ca/v-cats-kittens/markham-york-region/european-shorthair-kittens-for-adoption/16215051
5	1621504253	ontario	2022-06-15	850	british	na	short	female	Toronto	1009425732	3	https://kijiji.ca/v-cats-kittens/city-of-toronto/british-ahort-hair-kitten-for-rehome/1621504253
6	1621502937	ontario	2022-06-15	175	blue	blue	na	female	Ottawa	88111748	2	https://kijiji.ca/v-cats-kittens/ottawa/gorgeous-tri-coloured-female-kitten-for-sale/1621502937
7	1621502771	ontario	2022-06-15	250	shorthai	na	na	na	Nepean	1011049616	1	https://kijiji.ca/v-cats-kittens/ottawa/kittens-for-rehoming-they-are-so-adorable/1621502771
8	1621502764	ontario	2022-06-15	200	other	na	na	na	Milton	1024344666	1	https://kijiji.ca/v-cats-kittens/oakville-halton-region/adorable-kittens-for-sale/1621502764
9	1621502691	ontario	2022-06-15	500	bengal	orange	na	na	Toronto	1024898803	2	https://kijiji.ca/v-cats-kittens/city-of-toronto/gorgeous-orange-bengal-x-kittens-vaccinated/1621502691
10	1621502646	ontario	2022-06-15	400	ragdoll	na	long	female	Hamiltor	1005809334	1	https://kijiji.ca/v-cats-kittens/hamilton/long-haired-ragdoll-mix-kittens/1621502646
11	1621502036	ontario	2022-06-15	350	highland	grey	na	na	Toronto	1005478461	7	https://kijiji.ca/v-cats-kittens/city-of-toronto/purebred-highland-lynx-kittens-prices-reduced/16215020
12	1621501926	ontario	2022-06-15	200	other	na	na	na	Ottawa	1024867828	2	https://kijiji.ca/v-cats-kittens/ottawa/kitten/1621501926
13	1621501805	ontario	2022-06-15	750	ragdoll	na	na	male	Mississa	1004176252	4	https://kijiji.ca/v-cats-kittens/mississauga-peel-region/seal-bicolour-ragdoll-kitten-male/1621501805
14	1621501638	ontario	2022-06-15	175	other	na	na	male	Ottawa	88111748	2	https://kijiji.ca/v-cats-kittens/ottawa/2-beautiful-male-kittens-for-sale/1621501638
15	1621501141	ontario	2022-06-15	400	blue	blue	na	na	Toronto	55830552	8	https://kijiji.ca/v-cats-kittens/city-of-toronto/blue-russian-kitten/1621501141
16	1621500701	ontario	2022-06-15	100	other	black	short	male	Barrie	1009666059	7	https://kijiji.ca/v-cats-kittens/barrie/older-male-kittens-100-each-or-best-offer/1621500701
17	1621500492	ontario	2022-06-15	100	other	black	short	male	Kitchene	1009666059	7	https://kijiji.ca/v-cats-kittens/kitchener-waterloo/older-male-kittens-100-each-or-best-offer-black-whi
18	1621500313	ontario	2022-06-15	150	mixed	na	long	na	London	85088179	1	https://kijiji.ca/v-cats-kittens/london/long-haired-mixed-kittens/1621500313
19	1621500099	ontario	2022-06-15	100	other	black	short	male	Hamiltor	1009666059	7	https://kijiji.ca/v-cats-kittens/hamilton/older-male-kittens-100-each-or-best-offer-black-white-orange
20	1621500065	ontario	2022-06-15	275	other	na	na	female	Brampto	1012923606	1	https://kijiji.ca/v-cats-kittens/mississauga-peel-region/kitten-for-sale/1621500065
21	1621499893	ontario	2022-06-15	100	other	black	short	male	Markham	1009666059	7	https://kijiji.ca/v-cats-kittens/markham-york-region/older-male-kittens-100-each-or-best-offer-black-w
22	1621499851	ontario	2022-06-15	600	himalaya	blue	na	na	Markham	1024440833	1	https://kijiji.ca/v-cats-kittens/markham-york-region/himalayan-x-calico-kitten/1621499851
23	1621499800	ontario	2022-06-15	100	other	orange	na	female	Midland	1023074194	1	https://kijiji.ca/v-cats-kittens/barrie/female-orange-kitten-for-sale-100/1621499800

Photo credits: Kijiji

Data Analysis: Basic Statistics







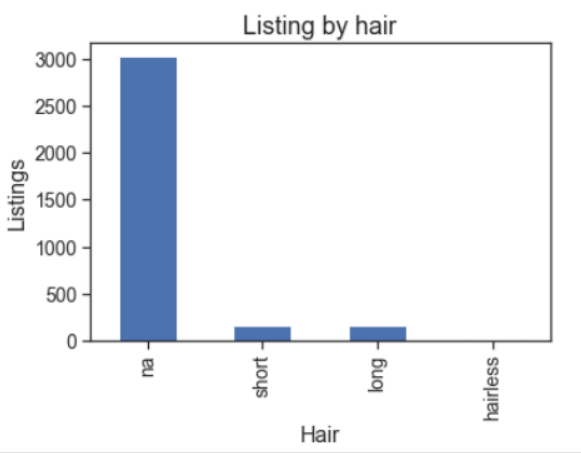
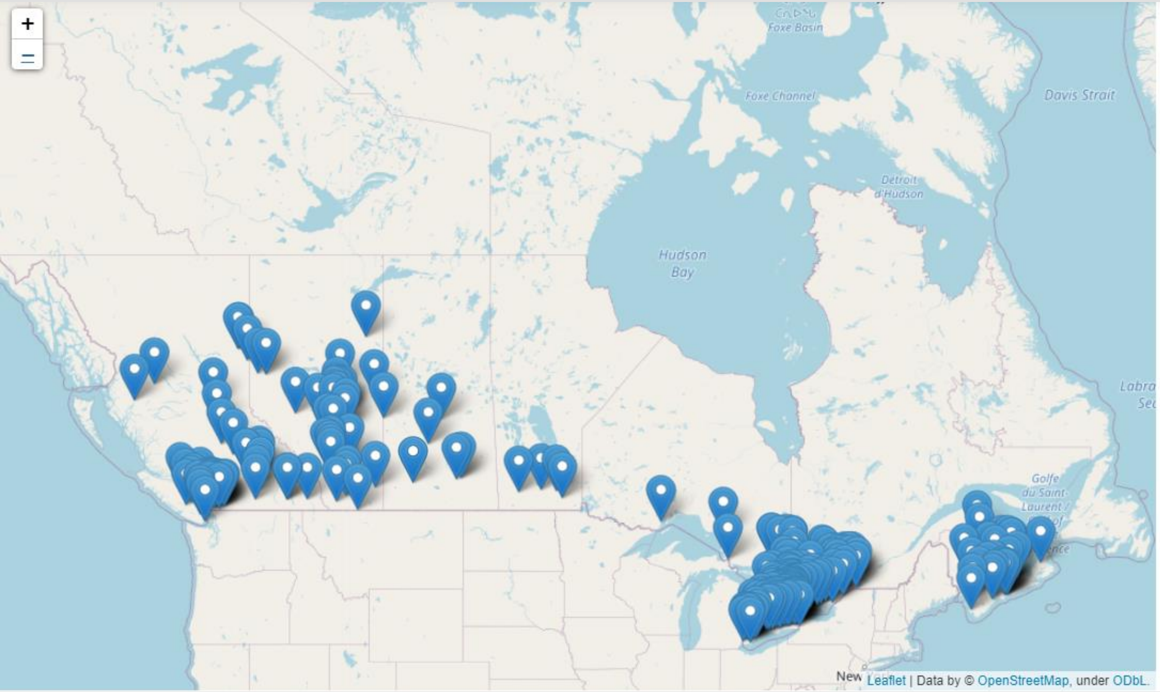


Photo credits: Kijiji

Data Analysis: Geo Distribution



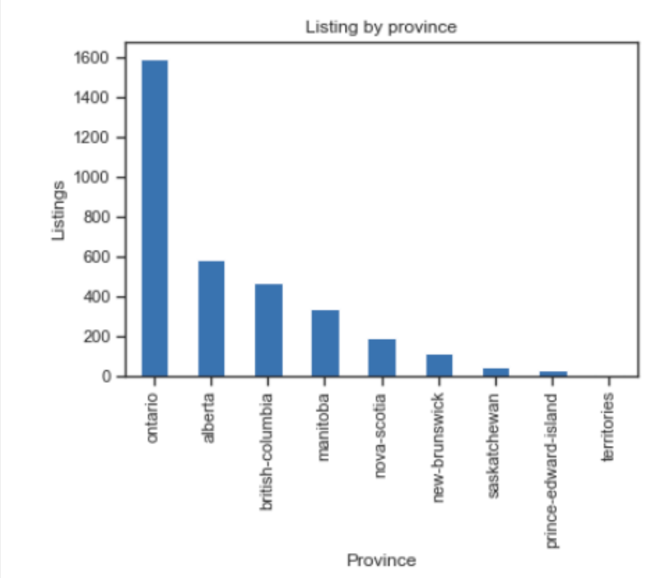
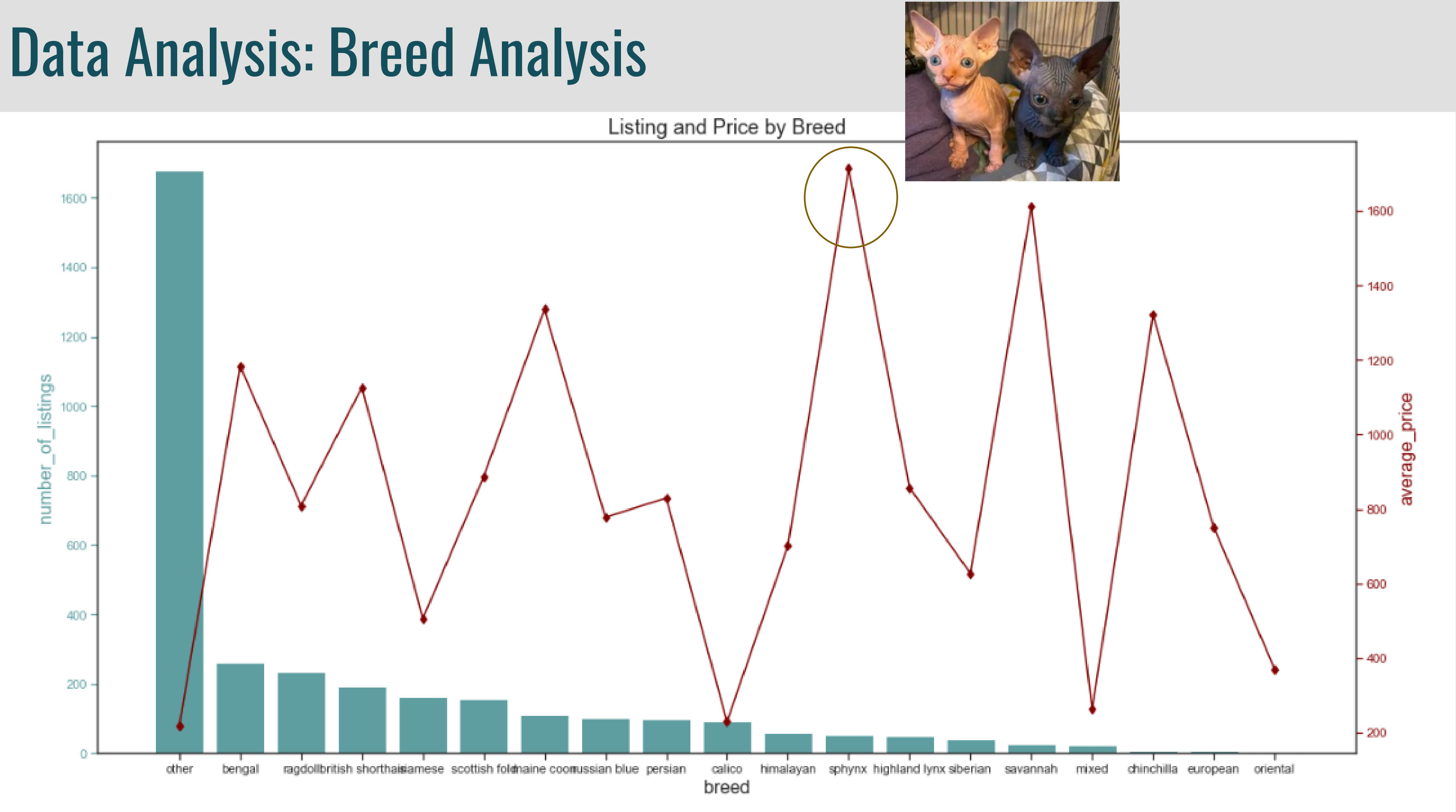


Photo credits: Kijiji

Data Analysis: Breed Analysis



Listing and Price by Breed

Photo credits: Kijiji

Data Analysis: Breed Analysis

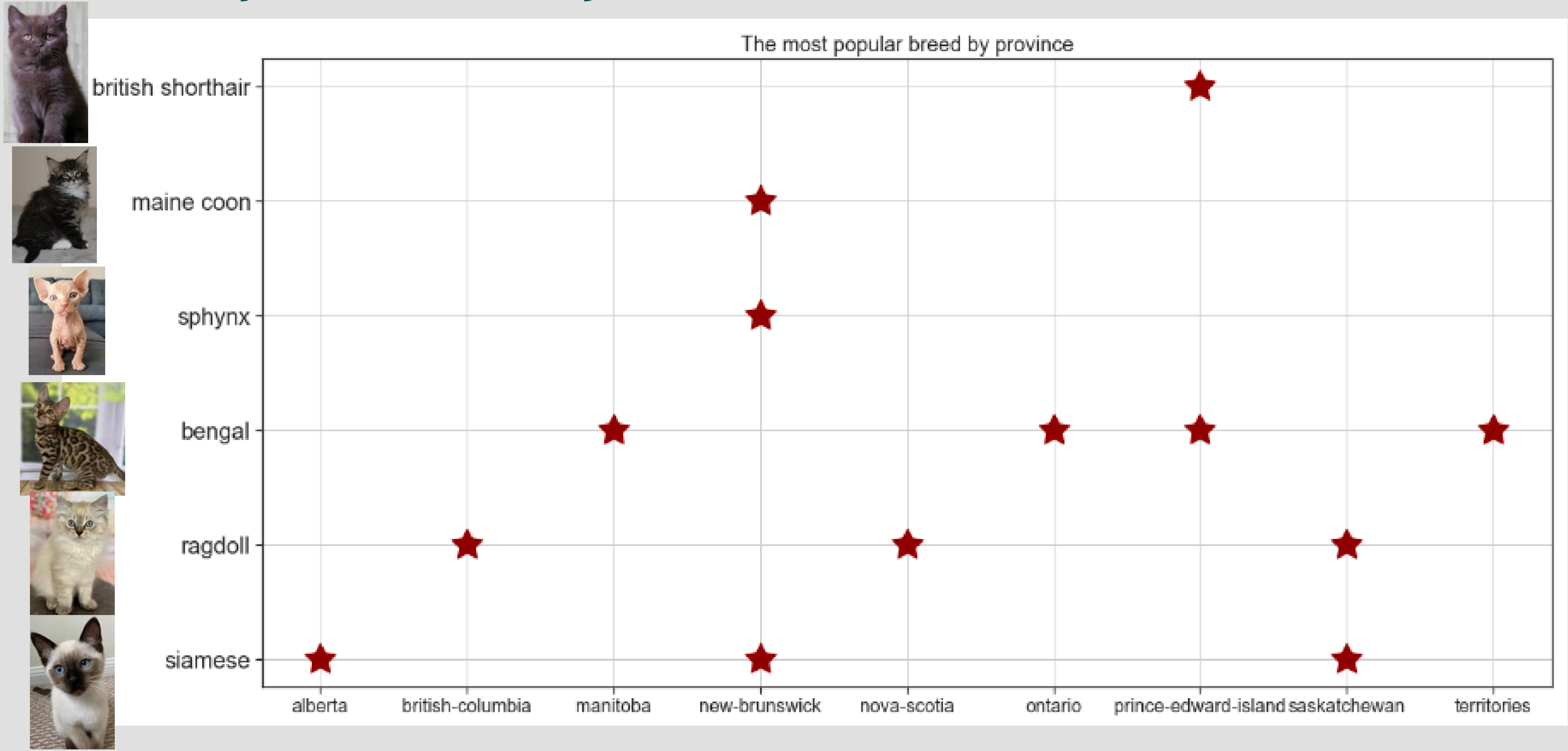
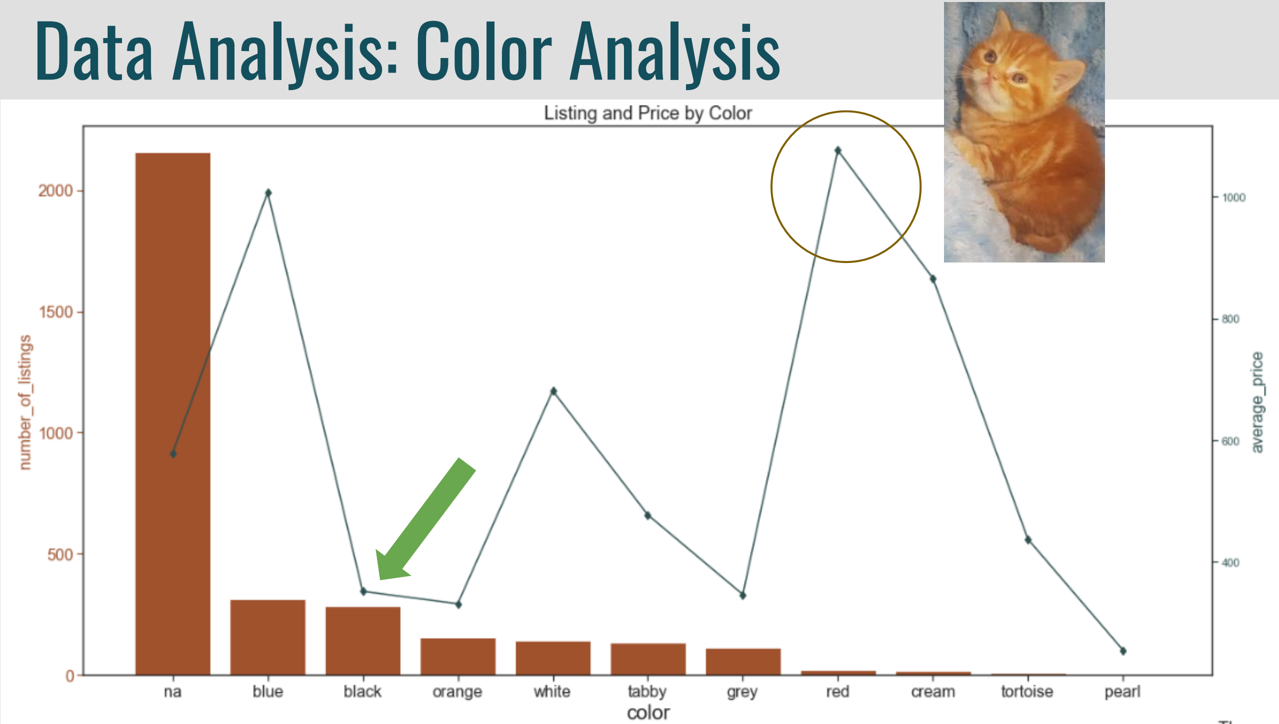


Photo credits: Kijiji

Note: excluding Others and Mix

Data Analysis: Color Analysis



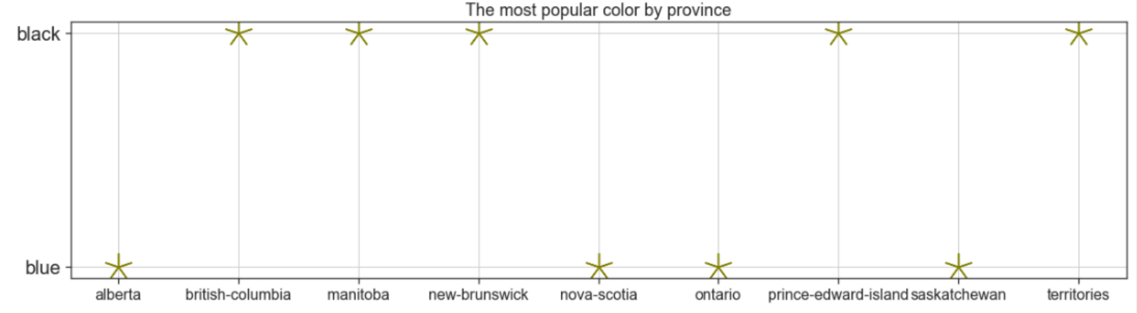


Photo credits: Kijiji

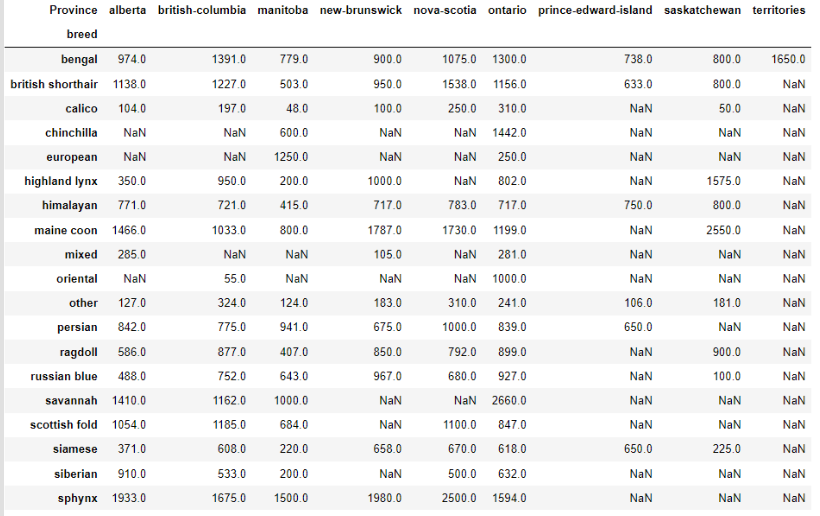
Interesting Findings and Conclusions

- Approximately half of the listings propose non-pedigree kittens

- Half of the listings accounts for Ontario

- Sphynx kittens are the most expensive across all data, however, in different provinces different breeds are the most expensive

- Black kittens are one of the cheapest compared to other colors, and that is one of the most popular colors



Province	alberta	british-columbia	manitoba	new-brunswick	nova-scotia	ontario	prince-edward-island	saskatchewan	territories
breed									
bengal	974.0	1391.0	779.0	900.0	1075.0	1300.0	738.0	800.0	1650.0
british shorthair	1138.0	1227.0	503.0	950.0	1538.0	1156.0	633.0	800.0	NaN
calico	104.0	197.0	48.0	100.0	250.0	310.0	NaN	50.0	NaN
chinchilla	NaN	NaN	600.0	NaN	NaN	1442.0	NaN	NaN	NaN
european	NaN	NaN	1250.0	NaN	NaN	250.0	NaN	NaN	NaN
highland lynx	350.0	950.0	200.0	1000.0	NaN	802.0	NaN	1575.0	NaN
himalayan	771.0	721.0	415.0	717.0	783.0	717.0	750.0	800.0	NaN
maine coon	1466.0	1033.0	800.0	1787.0	1730.0	1199.0	NaN	2550.0	NaN
mixed	285.0	NaN	NaN	105.0	NaN	281.0	NaN	NaN	NaN
oriental	NaN	55.0	NaN	NaN	NaN	1000.0	NaN	NaN	NaN
other	127.0	324.0	124.0	183.0	310.0	241.0	106.0	181.0	NaN
persian	842.0	775.0	941.0	675.0	1000.0	839.0	650.0	NaN	NaN
ragdoll	586.0	877.0	407.0	850.0	792.0	899.0	NaN	900.0	NaN
russian blue	488.0	752.0	643.0	967.0	680.0	927.0	NaN	100.0	NaN
savannah	1410.0	1162.0	1000.0	NaN	NaN	2660.0	NaN	NaN	NaN
scottish fold	1054.0	1185.0	684.0	NaN	1100.0	847.0	NaN	NaN	NaN
siamese	371.0	608.0	220.0	658.0	670.0	618.0	650.0	225.0	NaN
siberian	910.0	533.0	200.0	NaN	500.0	632.0	NaN	NaN	NaN
sphynx	1933.0	1675.0	1500.0	1980.0	2500.0	1594.0	NaN	NaN	NaN

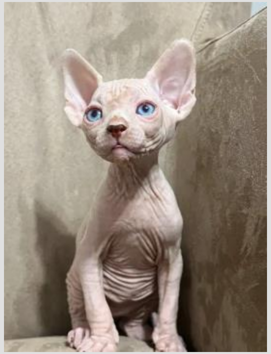


Photo credits: Kijiji

Challenges

- The scraping code took some time; the exact reason of mistake is not 100% clear and needs further investigation
- Kijiji does not provide separate fields for breed/color/gender, so this information has to be extracted from the plain text of title and/or description
- Data cleaning process was quite long
- Listings from Quebec should be processed separately (first, French; second, different price format)
- Data analysis using Pandas
- Getting latitude and longitude using Nominatim geocoder took some time
- Visualization using different libraries and functions

Photo credits: Kijiji

Next Steps

- Include listings from Quebec
- More meticulous analysis of the title and description (probably, url) to get breed/gender/color more accurate
- Seasonal analysis with long term time series data (seasons, holidays, etc.)
- Create a pricing model to predict price based on the location, breed, color, gender, age, etc.

Photo credits: Kijiji