Web Scraping

ZOLO.CA Real Estate Data Analyzing

Presenters:
Malik Aqib Rehman
&
Vivian Kuang

Date:
June 22nd, 2023

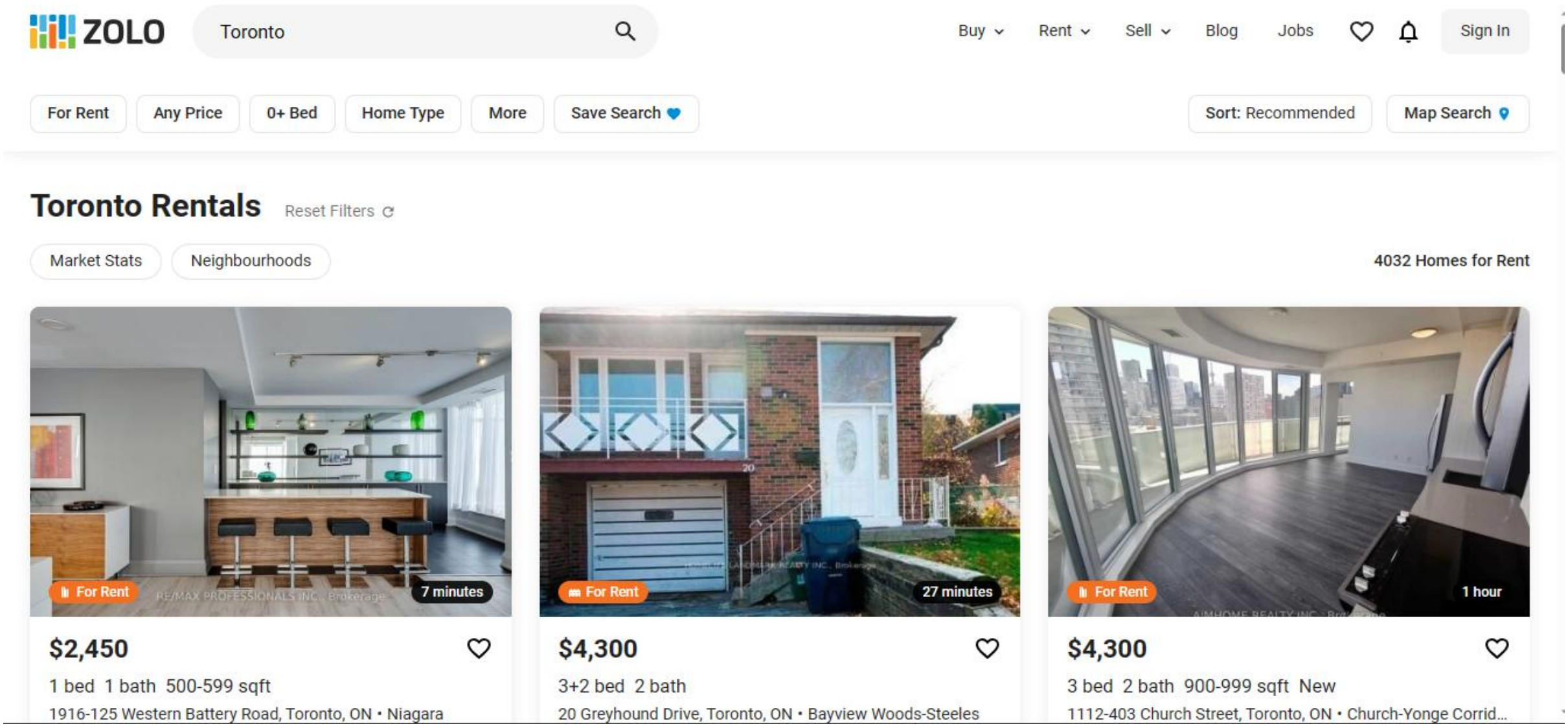
Contents

- Introduction
- Motivation
- Data Study and Analysis
- Conclusions and Challenges

Motivation

- Introducing Zolo.ca
- Motivation behind choosing Zolo.ca for web scraping:
  - Data Availability
  - Structured Website
  - Updated Listings

Website


ZOLO
Toronto
Buy
Rent
Sell
Blog
Jobs
Sign In
For Rent
Any Price
0+ Bed
Home Type
More
Save Search
Sort: Recommended
Map Search
Toronto Rentals
Reset Filters
Market Stats
Neighbourhoods
4032 Homes for Rent
For Rent
7 minutes
$2,450
1 bed 1 bath 500-599 sqft
1916-125 Western Battery Road, Toronto, ON • Niagara
For Rent
27 minutes
$4,300
3+2 bed 2 bath
20 Greyhound Drive, Toronto, ON • Bayview Woods-Steeles
For Rent
1 hour
$4,300
3 bed 2 bath 900-999 sqft New
1112-403 Church Street, Toronto, ON • Church-Yonge Corrid...

Dataframe

- Recorded data of 12709 listings for cities: Toronto, Vancouver, Surrey, Burnaby

URL	Address	City	Suburb	Beds	Baths	Size	Price	Type	Style	Taxes	Strata Fees	Walk Score	MLS ID	Listed By
/www.zolo.ca/vancouver-real-estate/2510...	2510 Fraser Street	Vancouver	Mount Pleasant Ve	3.00	3.00	1655.00	$1,250,000	Townhouse	4 Level Split	$4,191 /yr	$1,185 /mo	89	R2781630	MACDONALD REALTY
/www.zolo.ca/vancouver-real-estate/180-...	519 - 180 2nd Avenue E	Vancouver	Mount Pleasant Ve	1.00	1.00	598.00	$880,000	Apartment/Condo	Upper Unit	$1,995 /yr	$390 /mo	97	R2742811	RENNIE & ASSOCIATES REALTY LTD.
/www.zolo.ca/vancouver-real-estate/5955...	607 - 5955 Birney Avenue	Vancouver	University Vw	2.00	2.00	1033.00	$1,258,000	Apartment/Condo	Corner Unit	$1,508 /yr	$383 /mo	65	R2776687	UNILIFE REALTY INC.
/www.zolo.ca/vancouver-real-estate/63-w...	207 - 63 2nd Avenue W	Vancouver	False Creek	2.00	2.00	957.00	$1,188,000	Apartment/Condo	Upper Unit	$2,658 /yr	$645 /mo	94	R2789565	CENTURY 21 IN TOWN REALTY
/www.zolo.ca/vancouver-real-estate/161-...	108 - 161 King Edward Avenue W	Vancouver	Cambie	4.00	4.00	1837.00	$2,599,900	Townhouse	3 Storey	No Data	$569 /mo	65	R2777223	RENNIE & ASSOCIATES REALTY LTD.

Crawler

```python
cities = ['toronto']  # Add more cities as needed

# Create an empty dictionary to store the URLs for each listing
listings_urls = {}

for city in cities:
    for page_number in range(1, 171):  # Loop through pages 1 to 70
        # Generate the URL for the current city and page number
        url = f'https://www.zolo.ca/{city}-real-estate/page-{page_number}'

        # Send a GET request to the current page
        driver.get(url)
        page_source = driver.page_source
        soup = BeautifulSoup(page_source, 'html.parser')

        # Find the parent element by its ID
        parent_element = soup.find('section', id='gallery')

        # Find all elements within the parent element that have an 'href' attribute
        link_elements = parent_element.find_all(href=True)

        # Extract the URLs from the link elements and store them in the dictionary
        for link in link_elements:
            listing_id = link.get('data-listing-id')
            url = link['href']

            if listing_id in listings_urls:
                listings_urls[listing_id].add(url)
            else:
                listings_urls[listing_id] = {url}
```

```python
for listing_id, urls in listings_urls.items():
    for url in urls:
# for url in urls:
    # Open the URL in Selenium-controlled browser
        driver.get(url)

        # Find and extract the desired information using Selenium's find_element methods
        try:
            address_element = driver.find_element(By.CSS_SELECTOR, '#listing > div > div > section.xs-grid.xs-gap-2.listing-summa
            address = address_element.text.strip()
        except NoSuchElementException:
            address = ''

        try:
            country_element = driver.find_element(By.CSS_SELECTOR, '#listing > div > div > section.xs-grid.xs-gap-2.listing-summa
            city = country_element.text.strip()
        except NoSuchElementException:
            city = ''

        try:
            suburb_element = driver.find_element(By.CSS_SELECTOR, '#listing > div > div > section.xs-grid.xs-gap-2.listing-summar
            suburb = suburb_element.text.strip()
        except NoSuchElementException:
            suburb = ''

        try:
            beds_element = driver.find_element(By.CSS_SELECTOR, '#listing > div > div > section.xs-grid.xs-gap-2.listing-summary-
            beds = beds_element.text.strip()
        except NoSuchElementException:
            beds = ''
```

Processing the dataset

Clean the data

- Removing duplicates: Identifying and eliminating duplicate records
- Handling missing values: Addressing missing data by either imputing values or removing incomplete observations (Year Build, Walk Score)
- Correcting errors: Identifying and rectifying inaccurate or erroneous data entries (Price, Walk Score)

Transform the data

- Normalization: Scaling numerical data to a common range (e.g., between 0 and 1) to ensure equal importance.
- Standardization: Transforming data to have zero mean and unit variance, which helps when certain calculations

Select the features

- Selecting the most informative features for the specific analysis
- Choosing the most related labels regarding the business problems and modeling task

Analyzing & Visualization

- Calculating summary statistics such as mean, median, and standard deviation, to summarize the dataset's central tendencies and dispersion.
- Data profiling: Generating data summaries, frequency tables, and basic statistical analyses to gain initial insights into the data
- Creating bar plots, pie plots, scatter plots, histograms to show data

Interesting findings

Findings of the data

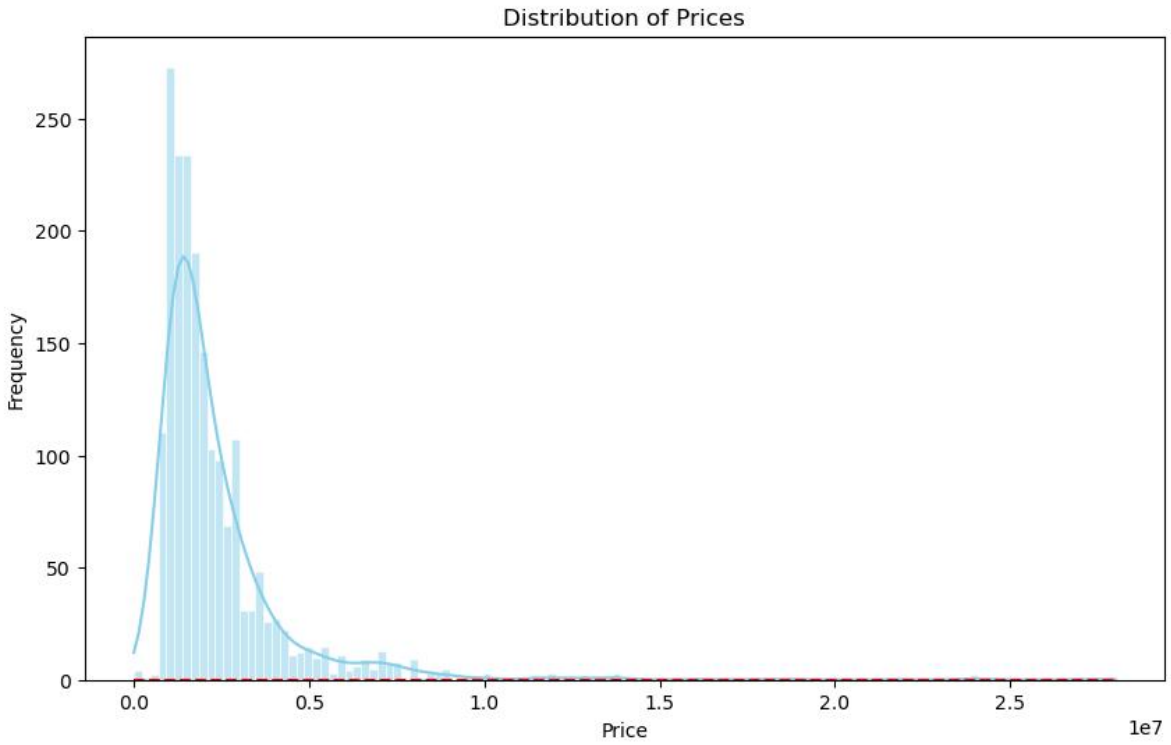
- Types (Condos/Apt, House/Townhouse) Walk Scores (1-100)
- Meaningful and Meaningless labels
- Price, lot size, bedrooms, bathrooms

```
RangeIndex: 2122 entries, 0 to 2121
Data columns (total 18 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   URL            2122 non-null   object
 1   Address        2121 non-null   object
 2   City           2121 non-null   object
 3   Suburb         2110 non-null   object
 4   Beds           1879 non-null   object
 5   Baths          1880 non-null   float64
 6   Size           663 non-null    object
 7   Offering Type  2120 non-null   object
 8   Price          1930 non-null   float64
 9   Type           2121 non-null   object
 10  Style          2121 non-null   object
 11  Lot Size       2121 non-null   object
 12  Year Built     55 non-null     object
 13  Walk Score     55 non-null     object
 14  MLS ID         1930 non-null   object
 15  Source         1930 non-null   object
 16  Listed By      1920 non-null   object
 17  final size     2044 non-null   float64
```

Interesting real estate

- Price Distribution
- Main Resource for the real estate data
- Types of properties for sale on the market
- 200 Prices change vs Neibourghhood, Type Correlation
- 50 highest-price vs lowest-price properties
- Prices change vs Lot Size Correlation
- Prices vs Bedrooms and Bathrooms Correlation

Analyzing the dataset of Toronto



	Baths	Price	final size
count	1880.000000	1.930000e+03	2044.000000
mean	3.604787	2.371595e+06	5968.915403
std	1.723172	2.150753e+06	7990.188459
min	1.000000	1.000000e+00	0.000000
25%	2.000000	1.249000e+06	2748.130200
50%	3.000000	1.749900e+06	4596.500000
75%	5.000000	2.695000e+06	6608.727950
max	19.000000	2.800000e+07	156711.766200

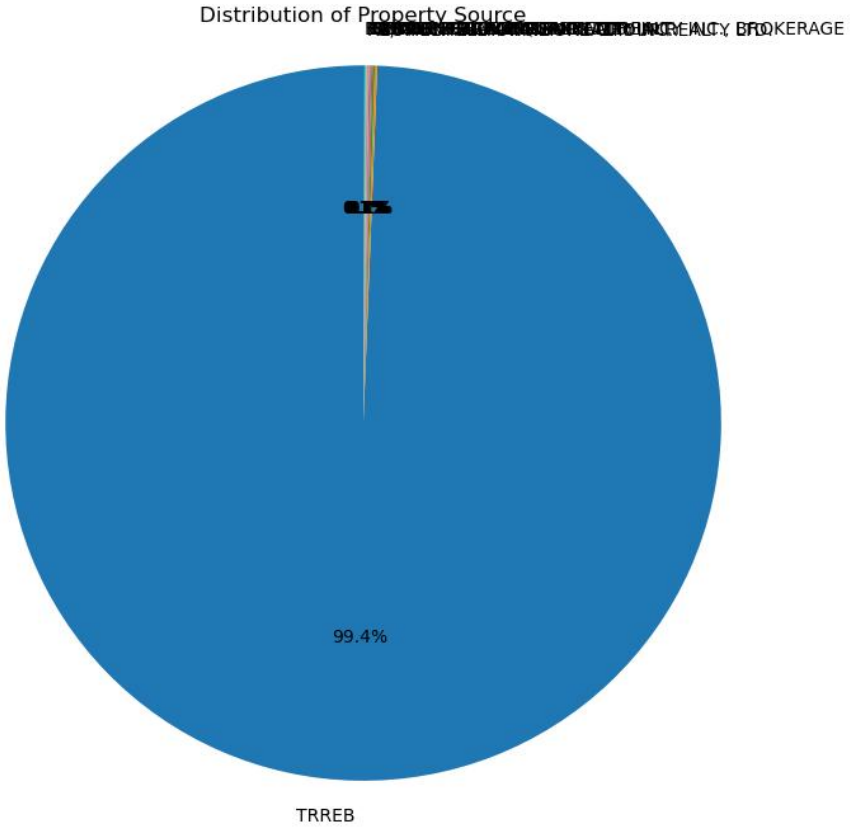
Insights

The frequency of listings' shows the most popular price of properties for sale

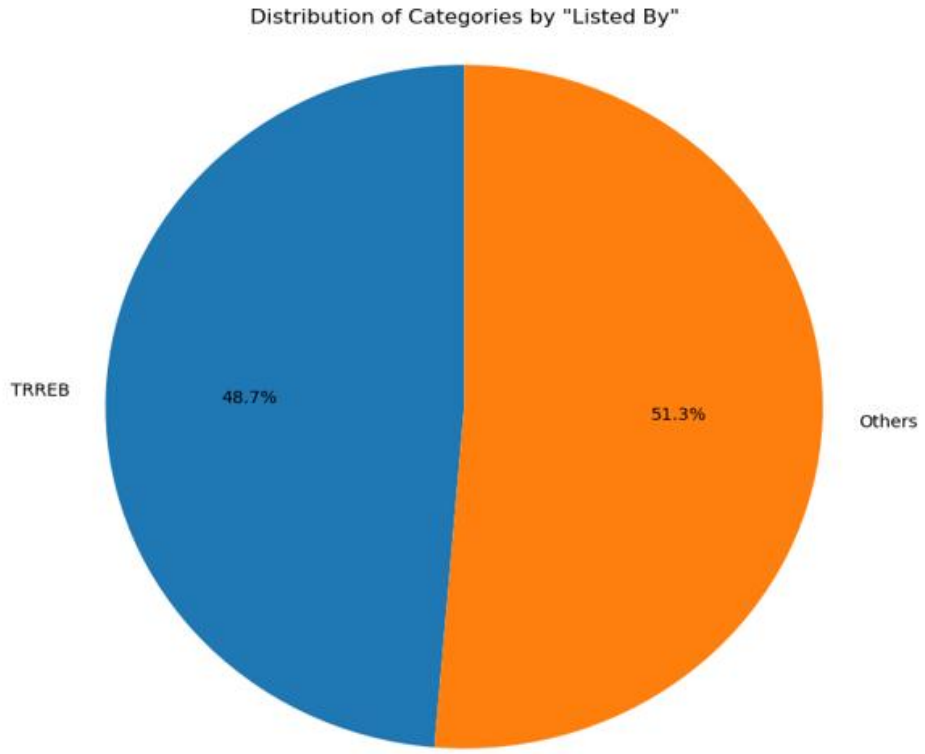
Insights

Statistic data for the three numeric labels of the dataset

Analyzing the dataset: Zoom In and Out

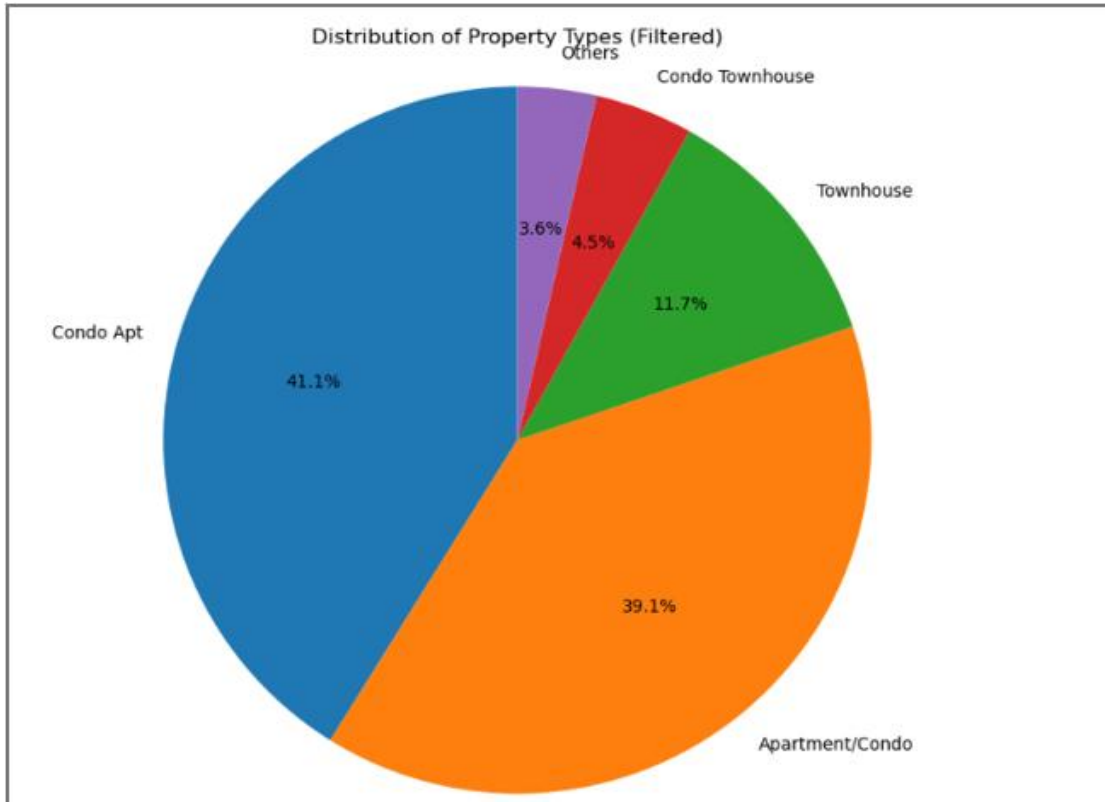


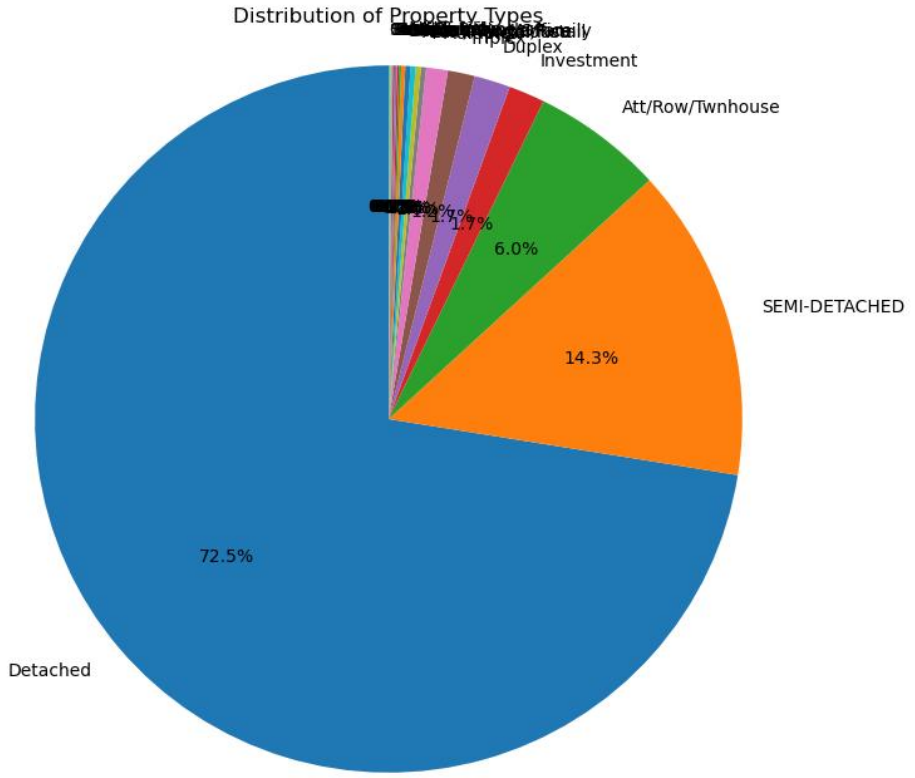
Pie Chart of Resource over Toronto data



Pie Chart of Resource over the 4-city data

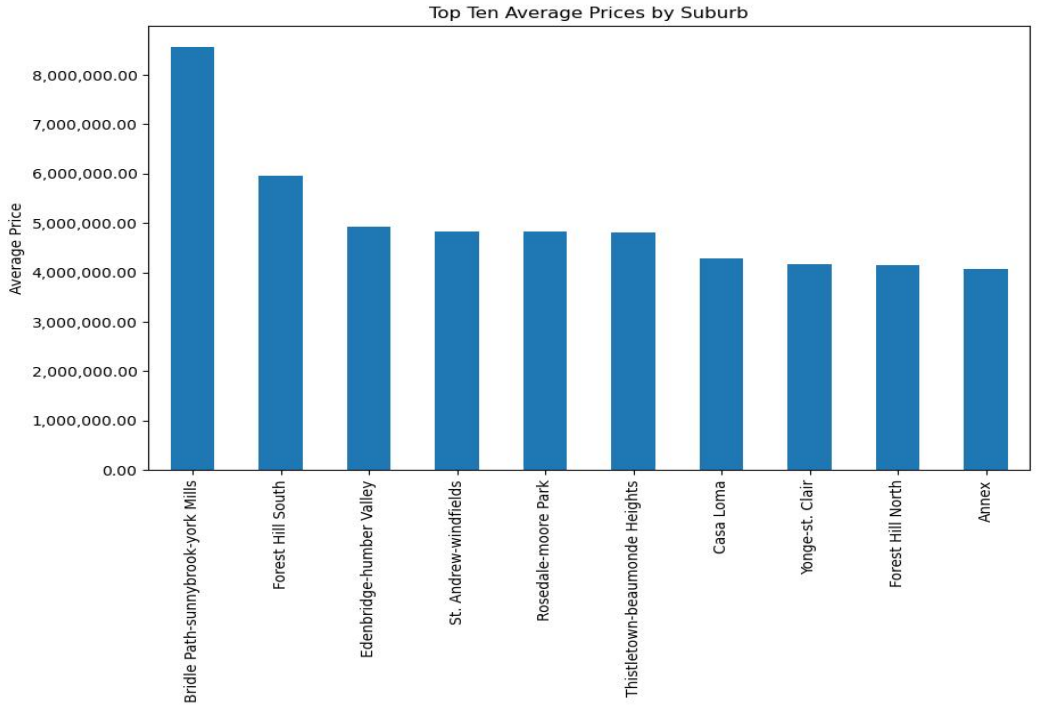
Analyzing the dataset

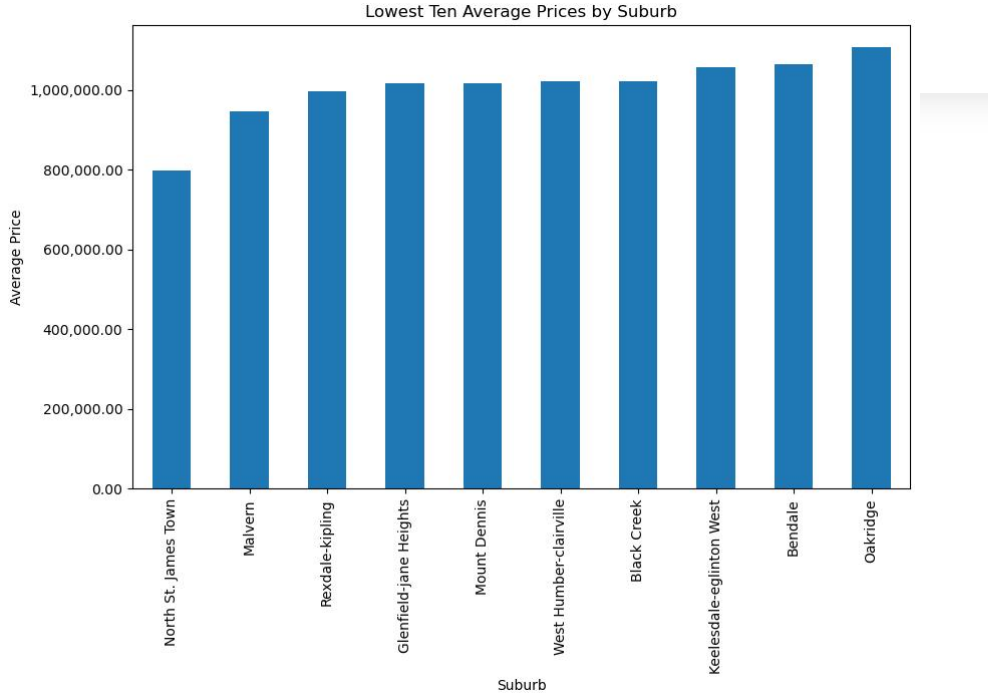


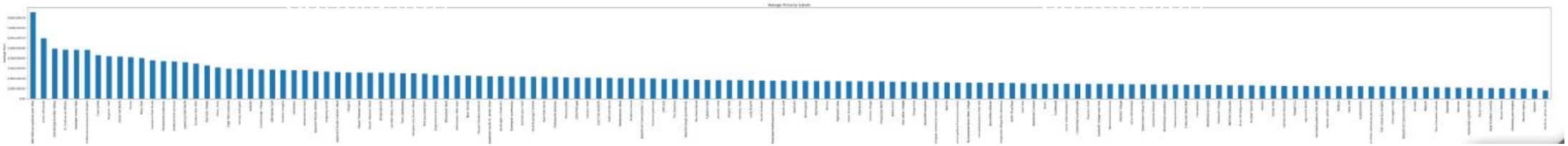


Pie Chart of types over properties Deep-In (ALL TYPES VS HOUSE TYPE DEEP-IN)

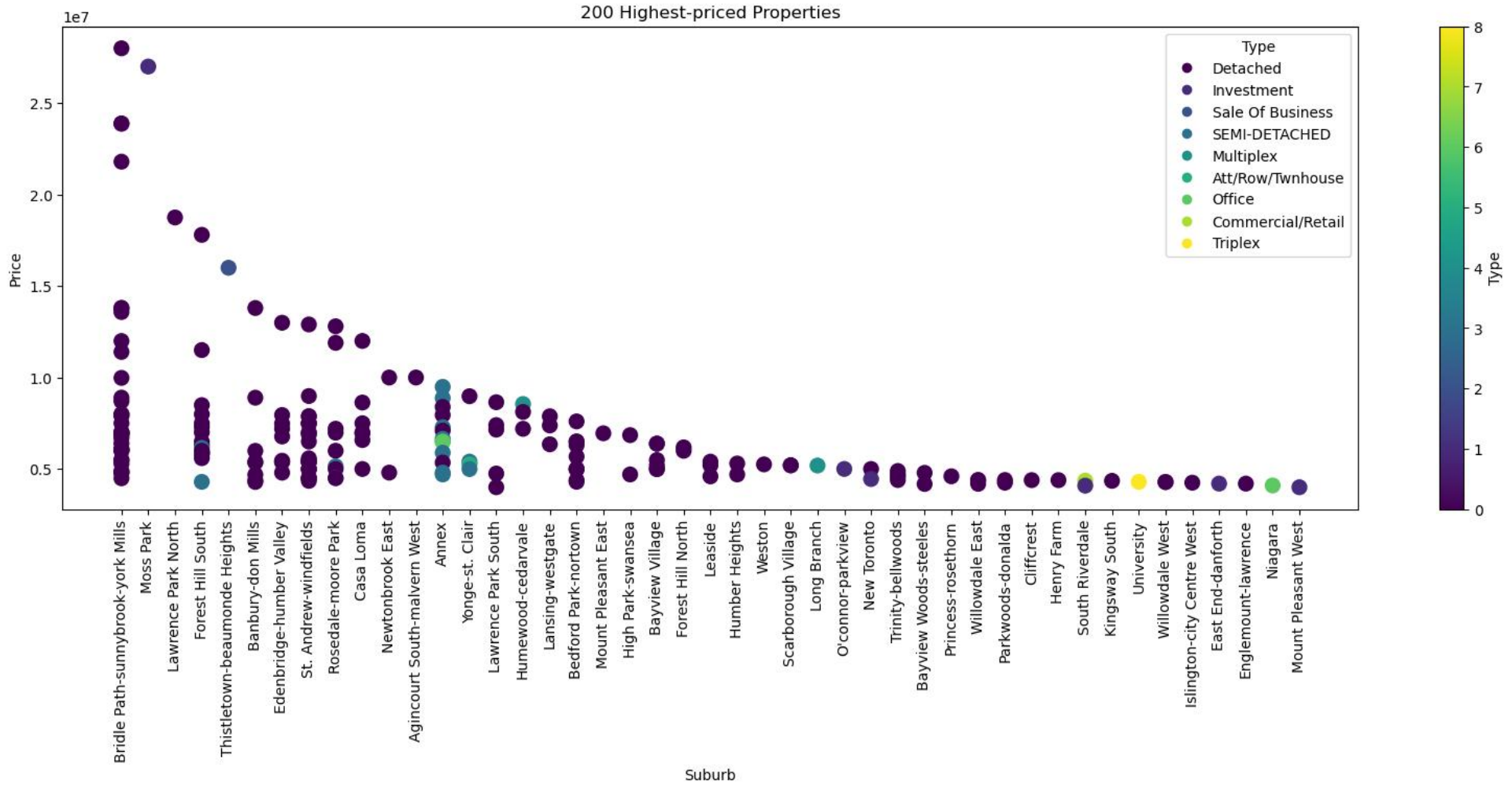
Analyzing the dataset – Business Sense







Analyzing the dataset – Bias vs Patterns



Conclusions

Conclusion

Enabling comprehensive data collection for market analysis, property valuation, and identifying investment opportunities

Empowering informed decision-making in the real estate industry.

Pricing and Sales Analysis

Property Trends and Market Analysis

Investment Opportunities

Challenges of the project

Web Scraping

1. Captchas and IP Blocking
   Locating the target labels from the HTML/Jason
2. Load chuck data

Data Processing

1. Data Consistency
2. Missing Data
3. Data Quality

Data Analyzing

1. Complex Data Relationships
2. Selection of Appropriate Analysis Techniques
3. Effective Visualization Design

Make the RIGHT recommendations/predictions/decisions

THANK YOU

QUESTIONS?

Presenters:
Vivian Kuang
&
Malik Aqib Rehman

Date:
June 22nd, 2023