



ETL Project

Real Estate/Walk Score Listings in Calgary



Team Members

- Alexis Lawal
- Dayo Thompson
- Kirushan Kirubaharan
- Sushant Deshpande



Objective

- Extracting Real Estate data in Calgary, Alberta and Walk Scores for corresponding house addresses
- Walk Score: shows a walk/transit/bike score (0-100) for any address to the local downtown
- Transforming retrieved data into easy-to-read tables
- Loading transformed data into relational and non-relational databases for optimal functionality

Data Sources

- Remax Canada: <https://www.remax.ca/ab/calgary-real-estate>
- Walk Score: <https://www.walkscore.com/CA-AB/Calgary>



Extract



Scraping Calgary Real Estate Data

Remax

```
In [1]: import pandas as pd
import numpy as np
import requests
from bs4 import BeautifulSoup
import time
from splinter import Browser
from sqlalchemy import create_engine
import warnings
warnings.filterwarnings('ignore')
print('Libraries imported!')
```

Libraries imported!

Using BeautifulSoup to scrape property details (house address, house details).

```
In [2]: house_address = []
house_details = []

base_url = 'https://www.remax.ca/ab/calgary-real-estate?page='
urls = [base_url + str(x) for x in range(1,301)]
time.sleep(2)
for url in urls:
    # Parse HTML with BeautifulSoup
    time.sleep(2)
    response = requests.get(url)
    soup = BeautifulSoup(response.text, 'html.parser')

    try:
        addresses = soup.find_all('div', class_='left-content flex-one')
        for address in addresses:
            house_address.append(address.text)
    except:
        house_address.append('None')

    try:
        details = soup.find_all('div', class_='property-details')
        for detail in details:
            house_details.append(detail.text)
    except:
        house_details.append('None')
```

Scraping Walk Score Data

```
In [ ]: scores_walk = []
scores_bike = []
scores_transit = []

for i in post_code_list:

    try:
        postal_code = i.replace(" ", "%20")
        url_score = "https://www.walkscore.com/score/" + str(postal_code)
        time.sleep(2)

        # Parse HTML with BeautifulSoup
        response = requests.get(url_score)
        code_soup = BeautifulSoup(response.text, 'html.parser')

        if 'pp.walk.sc/badge/walk/score' in str(code_soup):
            ws = str(code_soup).split('pp.walk.sc/badge/walk/score/')[1][:2].replace('.', '')
            scores_walk.append(ws)
        else:
            ws = 'N/A'
            scores_walk.append(ws)
        if 'pp.walk.sc/badge/bike/score' in str(code_soup):
            bs = str(code_soup).split('pp.walk.sc/badge/bike/score/')[1][:2].replace('.', '')
            scores_bike.append(bs)
        else:
            bs = 'N/A'
            scores_bike.append(bs)
        if 'pp.walk.sc/badge/transit/score' in str(code_soup):
            ts = str(code_soup).split('pp.walk.sc/badge/transit/score/')[1][:2].replace('.', '')
            scores_transit.append(ts)
        else:
            ts = 'N/A'
            scores_transit.append(ts)
    except:
        ws = 'N/A'
        scores_walk.append(ws)
        bs = 'N/A'
        scores_bike.append(bs)
        ts = 'N/A'
        scores_transit.append(ts)
```

Transform



Cleaning the Calgary Real Estate Data

First dataframe: Address and Price details

```
In [3]: address_df = pd.DataFrame(house_address)

new_df = address_df[0].str.split(' ', 2, expand=True)
new_df["price"] = new_df[1].str.replace("$", "")
new_df["price"] = new_df["price"].str.replace(",", "")
new_df["price"] = pd.to_numeric(new_df["price"])

del new_df[0]
del new_df[1]
new_df.head()
```

Out[3]:

	2	price
0	9803 ELBOW DR SW, Calgary, AB, T2V 1M4	489900
1	101 - 3704 15A ST SW, Calgary, AB, T2T 4C3	319900
2	25 HARVEST GLEN WAY NE, Calgary, AB, T3K 4J2	399900
3	32 EVERGLEN GROVE SW, Calgary, AB, T2Y 4Z3	429500
4	416 THORNDAL RD NW, Calgary, AB, T2K 3C5	484900

Cleaning the Calgary Real Estate Data

- Values separated into columns: price, address, postal code, bedrooms, bath, property type
- Price column type changed to integer

```
In [5]: final_df = new_df[2].str.split(', Calgary, AB, ', expand=True)
        final_df.head()
```

```
Out[5]:
```

	0	1
0	9803 ELBOW DR SW	T2V 1M4
1	101 - 3704 15A ST SW	T2T 4C3
2	25 HARVEST GLEN WAY NE	T3K 4J2
3	32 EVERGLEN GROVE SW	T2Y 4Z3
4	416 THORNDALE RD NW	T2K 3C5

```
In [6]: df_add = pd.concat([new_df, final_df], axis=1)
        del df_add[2]
        df_add.columns = ["price", "address", "postal_code"]
        df_add.head()
```

```
Out[6]:
```

	price	address	postal_code
0	489900	9803 ELBOW DR SW	T2V 1M4
1	319900	101 - 3704 15A ST SW	T2T 4C3
2	399900	25 HARVEST GLEN WAY NE	T3K 4J2
3	429500	32 EVERGLEN GROVE SW	T2Y 4Z3
4	484900	416 THORNDALE RD NW	T2K 3C5

Cleaning the Calgary Real Estate Data

- Second dataframe: House details

```
In [6]: details = pd.DataFrame(house_details)

details_df_temp = details[0].str.split('|', expand=True)

details_df_temp.head()
```

```
Out[6]:
```

	0	1	2	3
0	4 bed	2 bath	1121 sqft	house
1	2 bed	1 bath	836 sqft	condo
2	1 bed	1 bath	969 sqft	house
3	4 bed	2 bath	1650 sqft	house
4	4 bed	3 + 1 bath	2805 sqft	house

Joining the Calgary Real Estate Dataframes

- Concatenating House Address/Price details and House details dataframes.

```
In [12]: calgary_df_dup = pd.concat([df_add, details_df], axis=1)
calgary_df = calgary_df_dup.drop_duplicates()
calgary_df.head()
```

```
Out[12]:
```

	price	address	postal_code	bed	full_bath	half_bath	property_area	property_type
0	489900	9803 ELBOW DR SW	T2V 1M4	4.0	2.0	NaN	1121.0	house
1	239900	106 - 790 KINGSMERE CRES SW	T2V 2G9	2.0	1.0	NaN	836.0	condo
2	789900	4508 16A ST SW	T2T 4L7	1.0	1.0	NaN	969.0	house
3	595000	1015 19 AVE SE	T2G 1M1	4.0	2.0	NaN	1650.0	house
4	799900	96 ASPEN STONE RD SW	T3H 5Y7	4.0	3.0	1.0	2805.0	house

```
In [13]: calgary_df.to_csv('calgary_df.csv', index=False)
```

Cleaning the Walk Score Data

- Data converted into dataframe and columns named.

```
In [ ]: score_df_trans = {'postal_code':postal_code_list, 'walk_score':scores_walk, 'bike_score':scores_bike, 'transit_score':s  
score_df_dup = pd.DataFrame(score_df_trans)  
score_df = score_df_dup.drop_duplicates()  
score_df.head()
```

```
In [19]: score_df.to_csv('score_df.csv', index=False)
```

```
In [9]: score_df.head()
```

```
Out[9]:
```

	postal_code	walk_score	bike_score	transit_score
0	T2V 1M4	58.0	61.0	55.0
1	T2T 4C3	53.0	81.0	42.0
2	T3K 4J2	19.0	59.0	38.0
3	T2Y 4Z3	6.0	30.0	31.0
4	T2K 3C5	61.0	81.0	53.0

Load



Loading Data to Relational Database (PostgreSQL)

- Building connection to PostgreSQL and loading transformed data

SQL

```
In [29]: calgary_df = pd.read_csv('calgary_df.csv')  
score_df = pd.read_csv('score_df.csv')
```

```
In [30]: rds_connection_string = "postgres:1@localhost:5432/realestate_db"  
engine = create_engine(f'postgresql://{rds_connection_string}')  
  
calgary_df.to_sql(name= "calgary_df", con=engine, if_exists="append", index=False)  
score_df.to_sql(name= "score_df", con=engine, if_exists="append", index=False)
```

PostgreSQL Database

pgAdmin File Object Tools Help

Browser Dashboard Properties SQL Statistics Dependencies Dependents public.calgary_df/realestate_db/postgres@PostgreSQL 12

100 rows

public.calgary_df/realestate_db/postgres@PostgreSQL 12

Query Editor Query History Scratch Pad

```
1 SELECT * FROM public.calgary_df
2 LIMIT 100
3
```

Data Output Explain Messages Notifications

	price integer	address text	postal_code text	bedrooms text	bath text	property_type text
1	489900	9803 ELBOW...	T2V 1M4	4 bed	2 bath	house
2	319900	101 - 3704 1...	T2T 4C3	2 bed	1 + 1 bath	condo
3	399900	25 HARVEST ...	T3K 4J2	3 bed	2 bath	house
4	429500	32 EVERGLE...	T2Y 4Z3	3 bed	2 + 1 bath	house
5	484900	416 THORND...	T2K 3C5	3 bed	2 bath	house
6	529000	46 HARVEST ...	T3K 4T6	4 bed	3 + 1 bath	house
7	379900	26 ANAHEIM...	T1Y 7B3	4 bed	2 + 1 bath	house
8	337000	1905 - 930 6 ...	T2P 1J3	1 bed	1 bath	condo
9	544500	52 RIVERVIE...	T2C 3Z8	4 bed	3 bath	house
10	445000	1 - 435 13 AV...	T2E 1C3	3 bed	2 + 1 bath	townhouse

PostgreSQL Database

The screenshot displays the PgAdmin 4 web interface. The left sidebar shows a tree view of database objects, with the 'public' schema selected. The main pane shows the 'Query Editor' with a SQL query: `SELECT * FROM public.calgary_df LIMIT 100`. Below the query editor, the 'Data Output' tab is active, displaying a table with 12 rows of data. The table columns are: price (bigint), address (text), postal_code (text), bed (double precision), full_bath (double precision), half_bath (double precision), property_area (double precision), and property_type (text).

	price bigint	address text	postal_code text	bed double precision	full_bath double precision	half_bath double precision	property_area double precision	property_type text
1	489900	9803 ELBOW...	T2V 1M4		4	2	[null]	1121 house
2	239900	106 - 790 KIN...	T2V 2G9		2	1	[null]	836 condo
3	789900	4508 16A ST ...	T2T 4L7		1	1	[null]	969 house
4	595000	1015 19 AVE ...	T2G 1M1		4	2	[null]	1650 house
5	799900	96 ASPEN ST...	T3H 5Y7		4	3	1	2805 house
6	449900	12967 COVE...	T3K 5R2		3	3	1	1730 house
7	469000	7610 MARTH...	T3J 3X7		4	3	1	1785 house
8	379900	55 EVERGLE...	T2Y 4Z4		3	1	1	1416 house
9	139500	9206 - 315 S...	T2W 2T6		2	1	[null]	815 condo
10	955000	31 LYNX ME...	T3L 2M1		5	3	1	3006 house
11	184900	406 - 314 14 ...	T2N 1Z8		1	1	[null]	572 condo
12	349900	206 LYNNWO...	T2C 0S9		4	2	[null]	1195 house

Loading Data to Non-Relational Database (MongoDB)

- Building connection to MongoDB and loading transformed data

MongoDB

```
In [ ]: # Make a connection
conn = "mongodb://localhost:27017"

# Making a Connection with MongoClient
client = MongoClient(conn)

# database
db = client.realestate_db

collection = db.calgary
calgary_dict = calgary_df.to_dict("records")
collection.insert_many(calgary_dict)

collection = db.score
score_dict = score_df.to_dict("records")
collection.insert_many(score_dict)
```

MongoDB

Connect View Collection Help

Local

8 DBS 7 COLLECTIONS

☆ FAVORITE

HOST
localhost:27017

CLUSTER
Standalone

EDITION
MongoDB 4.4.0 Community

Filter your data

- > Dumpster_DB
- > admin
- > config
- > fruits_db
- > local
- > mars_db
- ▼ realestate_db
 - calgary_df_html ...
 - score_df_html
 - > store_inventory

realestate_db.calgary_df_html Documents

Documents 2.8k TOTAL SIZE 514.6KB AVG. SIZE 191B INDEXES 1 TOTAL SIZE 40.0KB AVG. SIZE 40.0KB

Documents Aggregations Explain Plan Indexes

FILTER OPTIONS FIND RESET

ADD DATA VIEW

Displaying documents 1 - 20 of 2752 REFRESH

```
{
  "_id": ObjectId("5f3b65c49bab9561bf7c7db0"),
  "price": 489900,
  "address": "9803 ELBOW DR SW",
  "postal_code": "T2V 1M4",
  "bed": 4,
  "full_bath": 2,
  "half_bath": NaN,
  "property_area": 1121,
  "property_type": "house"
}
```

```
{
  "_id": ObjectId("5f3b65c49bab9561bf7c7db1"),
  "price": 239900,
  "address": "106 - 790 KINGSMERE CRES SW",
  "postal_code": "T2V 2G9",
  "bed": 2,
  "full_bath": 1,
  "half_bath": NaN,
  "property_area": 836,
  "property_type": "condo"
}
```

```
{
  "_id": ObjectId("5f3b65c49bab9561bf7c7db2"),
  "price": 789900,
  "address": "4508 16A ST SW",
  "postal_code": "T2T 4L7",
  "bed": 1,
  "full_bath": 1,
  "half_bath": NaN,
  "property_area": 969,
  "property_type": "house"
}
```

Converted Database into a Web Based Application

Remax Calgary Real Estate					
Price (\$)	Address	Postal Code	Bedrooms	Bathrooms	Property Type
489900	9803 ELBOW DR SW	T2V 1M4	4 bed	2 bath	house
319900	101 - 3704 15A ST SW	T2T 4C3	2 bed	1 + 1 bath	condo
399900	25 HARVEST GLEN WAY NE	T3K 4J2	3 bed	2 bath	house
429500	32 EVERGLEN GROVE SW	T2Y 4Z3	3 bed	2 + 1 bath	house
484900	416 THORNDALE RD NW	T2K 3C5	3 bed	2 bath	house
529000	46 HARVEST GROVE CLOSE NE	T3K 4T6	4 bed	3 + 1 bath	house
379900	26 ANAHEIM PL NE	T1Y 7B3	4 bed	2 + 1 bath	house
337000	1905 - 930 6 AVE SW	T2P 1J3	1 bed	1 bath	condo
544500	52 RIVERVIEW MEWS SE	T2C 3Z8	4 bed	3 bath	house
445000	1 - 435 13 AVE NE	T2E 1C3	3 bed	2 + 1 bath	townhouse
229000	3 - 203 VILLAGE TERR SW	T3H 2L4	2 bed	2 bath	condo
409900	234 ROYAL BIRCH BAY NW	T3G 5X6	3 bed	3 bath	house
749900	111 HILLGROVE CRES SW	T2V 3K9	2 bed	1 bath	house
924900	3030 26A ST SW	T3E 2E3	3 bed	2 bath	house
574900	78 CHAPARRAL VALLEY GROVE SE	T2X 0M4	4 bed	3 + 1 bath	house
438700	203 ARBOUR STONE PL NW	T3G 5E9	3 bed	2 + 1 bath	house
369900	108 - 59 22 AVE SW	T2S 3C7	2 bed	2 bath	condo
489000	140 CITADEL CREST CIR NW	T3G 4G3	3 bed	2 + 1 bath	house
469900	25 DOUGLASBANK RISE SE	T2Z 2C5	4 bed	2 + 1 bath	house
Postal Code	Walk Score	Bike Score	Transit Score		
T2V 1M4	58	61	55		
T2T 4C3	53	81	42		
T3K 4J2	19	59	38		
T2Y 4Z3	6	30	31		
T2K 3C5	61	81	53		
T3K 4T6	40	60	39		
T1Y 7B3	13	59	40		
T2P 1J3	94	94	78		
T2C 3Z8	32	70	39		
T2E 1C3	82	84	54		
T3H 2L4	26	67	35		
T3G 5X6	47	65	36		
T2V 3K9	36	53	53		
T3E 2E3	56	74	48		
T2X 0M4	4	50	28		
T3G 5E9	15	47	49		
T2S 3C7	55	91	59		
T3G 4G3	23	65	38		