Real Estate Analysis

By Rachael Fisher and Andy Ward

Data collection

- Python script to scrape data from Zillow

- BeautifulSoup or Scrapy?

- CAPTCHAS

- Headless browser by using Selenium

- 181 results

- CSV

Data Collection continued

```python
for page in range(1, 21):
    url = f'https://www.zillow.com/ut/{page}_p/'
    data = requests.get(url, headers=header)
    soup = BeautifulSoup(data.text, 'lxml')

    for result in soup.find_all('ul', {'class': 'dmDolk'}):
        li_list = result.find_all('li')
        if len(li_list) > 1:
            sqr.append(li_list[2].text)
        else:
            sqr.append('N/A')
    # Address
    for result in soup.find_all('address', {'data-test':'property-card-addr'}):
        adr.append(result.text)
    # Price
    for result in soup.find_all('span', {'data-test':'property-card-price'}):
        pr.append(result.text)
```

Web Scraping for National Park/Monument data and LDS Temples Location data

```
,name,Latitude,Longitude
0,Arches National Park,38.72261844,-109.5863666
1,Bryce Canyon National Park,37.58399144,-112.1826689
2,California National Historic Trail,42.3999643979,-108.702415369
3,Canyonlands National Park,38.24555783,-109.8801624
4,Capitol Reef National Park,38.2821653131,-111.247048378
5,Cedar Breaks National Monument,37.63545562,-112.845268
6,Dinosaur National Monument,40.50739742,-108.9333714
7,Glen Canyon National Recreation Area,36.9357464677,-111.485594268
8,Golden Spike National Historical Park,41.61947508,-112.5242653
9,Hovenweep National Monument,37.39134092,-109.0047977
10,Mormon Pioneer National Historic Trail,41.2650321741,-101.84083756
11,Natural Bridges National Monument,37.6045329,-110.0022436
12,Old Spanish National Historic Trail,37.0791782514,-112.114693431
13,Pony Express National Historic Trail,42.2365761128,-109.266148417
14,Rainbow Bridge National Monument,36.9943264317,-111.491484759
15,Timpanogos Cave National Monument,40.44039719,-111.7093944
16,Zion National Park,37.29839254,-113.0265138
```

```
templeId,address,latitude,longitude
0,bountiful-utah-temple,640 S Bountiful Blvd Bountiful UT  84010-1394 United States,,
1,brigham-city-utah-temple,250 S Main St Brigham City UT  84302-2560 United States,,
2,cedar-city-utah-temple,280 South Cove Dr Cedar City UT  84720 United States,,
3,deseret-peak-utah-temple,2400 N 400 W Tooele UT 84074 United States,,
4,draper-utah-temple,14065 Canyon Vista Ln Draper UT  84020-5626 United States,,
5,jordan-river-utah-temple,10200 S Temple Drive South Jordan UT  84095-8814 United States,,
6,layton-utah-temple,1400 Oak Hills Dr Layton UT 84040 United States,,
7,logan-utah-temple,175 N 300 E Logan UT  84321-4720 United States,,
8,manti-utah-temple,200 E 510 N Manti UT  84642-1701 United States,,
9,monticello-utah-temple,365 N 200 W Monticello UT  84535 United States,,
10,mount-timpanogos-utah-temple,742 N 900 E American Fork UT  84003-9124 United States,,
11,ogden-utah-temple,350 22nd St Ogden UT  84401-1487 United States,,
12,oquirrh-mountain-utah-temple,11022 S 4000 W South Jordan UT  84009-5797 United States,,
13,orem-utah-temple,1471 S Geneva Rd Orem UT  84058 United States,,
14,payson-utah-temple,1494 S 930 W Payson UT  84651 United States,,
15,provo-city-center-temple,50 South University Ave Provo UT  84601 United States,,
16,provo-utah-temple,2200 Temple Hill Dr Provo UT  84604-1775 United States,,
17,red-cliffs-utah-temple,1580 S St St George UT 84790 United States,,
18,salt-lake-temple,50 W North Temple St Salt Lake City UT  84150-9709 United States,,
19,saratoga-springs-utah-temple,987 South Ensign Drive Saratoga Springs UT  84045-3839 United
20,st.-george-utah-temple,250 E 400 S St George UT  84770-3699 United States,,
21,syracuse-utah-temple,1025 S 2500 W Syracuse UT 84075 United States,,
22,taylorsville-utah-temple,2603 W 4700 S Taylorsville UT  84129 United States,,
23,vernal-utah-temple,420 W 200 S Vernal UT  84078-2536 United States,,
```

Data Labeling

Addresses, Price, Square Feet, Latitude, Longitude

templeId,address,latitude,longitude

name,Latitude,Longitude

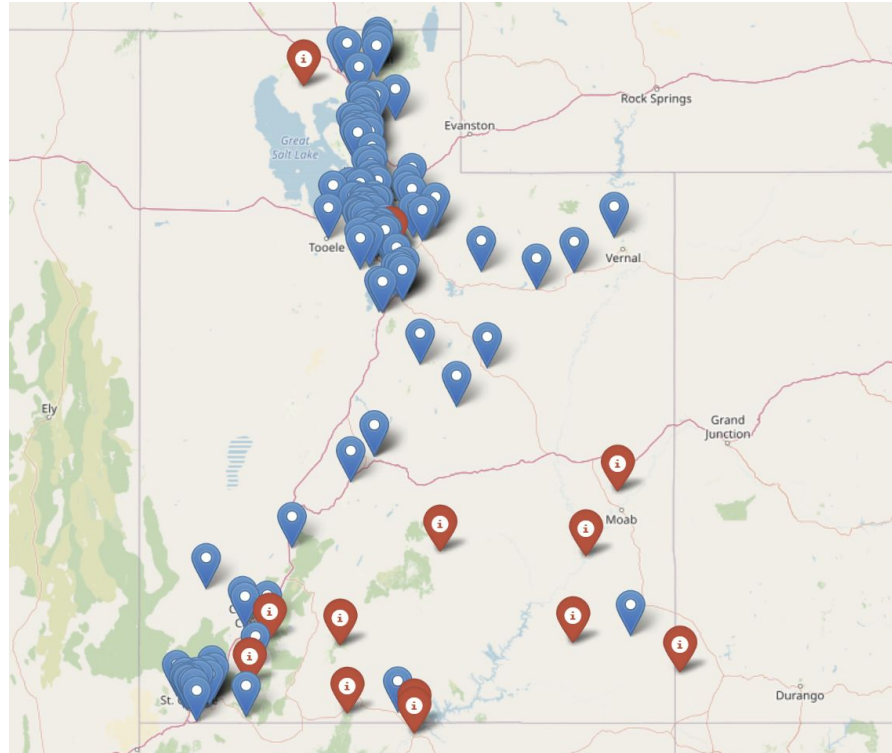## Data Preprocessing

Cleaning: Drop N/A

```
zillow_df = zillow_df.drop(zillow_df[zillow_df['Address'].isin(['Canyon Plan, Juniper Trail', 'Stanton
Plan, Juniper Trail', 'Cambria Plan, Juniper Trail, N Highway 24 Hwy, Hanksville, UT 84734'])].index)
```

- Strip , $ sqrft

```
parks_df = parks_df.drop(parks_df[parks_df['name'] == 'California National Historic Trail'].index)
parks_df = parks_df.drop(parks_df[parks_df['name'] == 'Pony Express National Historic Trail'].index)
```

```
zillow_df['price_per_sqft'] = zillow_df['Price'] / zillow_df['Square Feet']
```
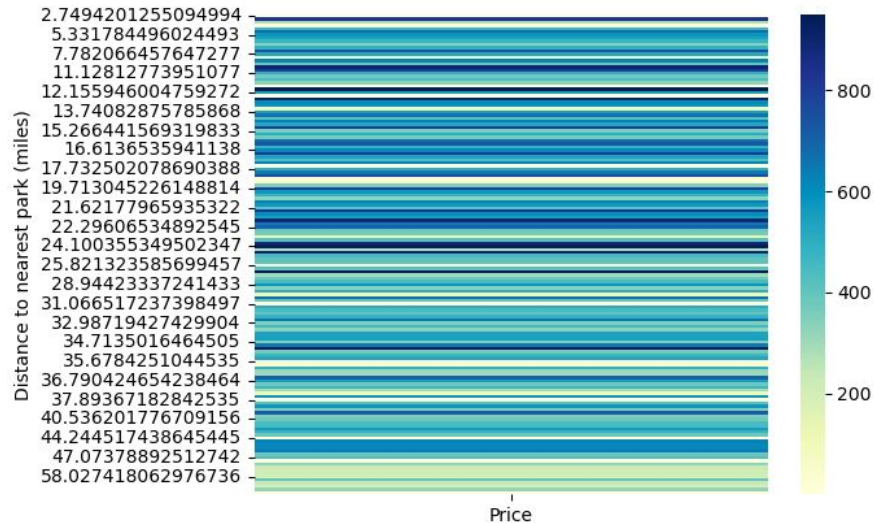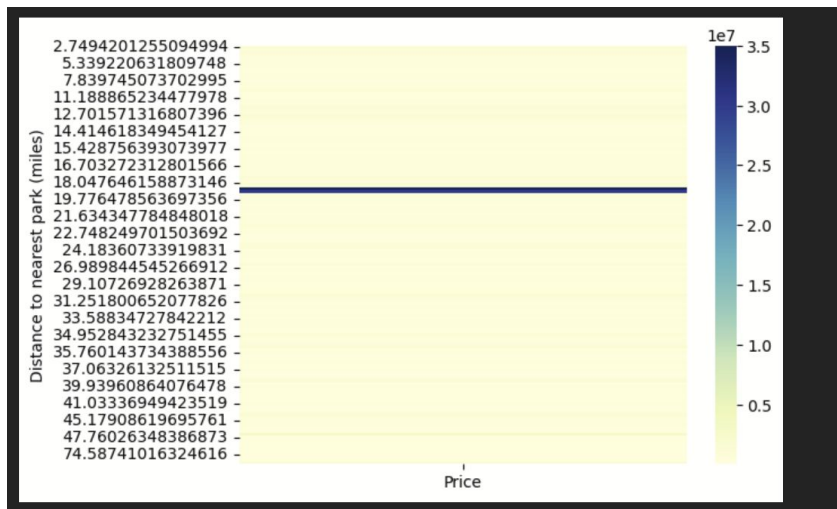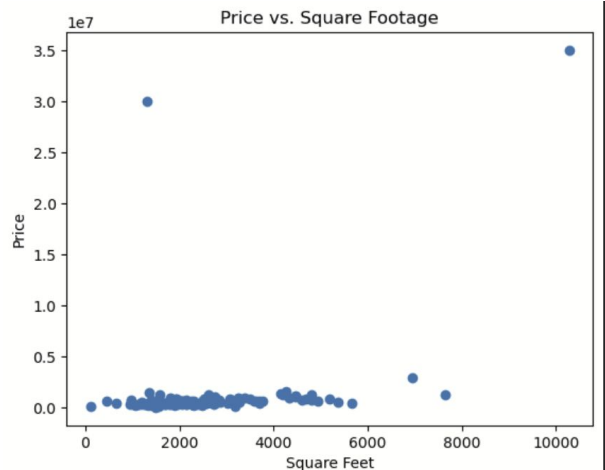
Folium - web maps for your data

Modeling

Results

Analysis

Unnamed: 0	Address	Price	Square Feet	Latitude	Longitude	price_per_sqft
49	2470 W White Pine Ln, Park City, UT 84060	35000000.0	10286.0	40.678579	-111.558335	3402.683259
170	2189 W White Pine Canyon Rd, Park City, UT 84060	30000000.0	1308.0	40.676847	-111.552769	22935.779817
114	3519 N Layton Ridge Dr, Layton, UT 84040	2900000.0	6955.0	41.111094	-111.904507	416.966211
37	136 N 800 E, Genola, UT 84655	1595000.0	4266.0	39.999336	-111.823811	373.886545
147	833 E Mountain Rd, Ogden, UT 84414	1499000.0	1350.0	41.329014	-111.954438	1110.370370
...	...	...	...	...	...	...
141	3394 S Westcrest Rd #18, West Valley City, UT ...	120000.0	3176.0	40.698751	-111.992660	37.783375
128	1031 N 500 E, Ogden, UT 84404	110000.0	1560.0	41.276220	-111.966432	70.512821
45	2698 E Cedar Dr, Virgin, UT 84779	99500.0	115.0	37.404381	-112.975555	865.217391
104	780 W 1125 N UNIT 50, Cedar City, UT 84721	95000.0	1500.0	37.697339	-113.074891	63.333333
157	350 W 800 S UNIT 53, Logan, UT 84321	34999.0	1495.0	41.716729	-111.844638	23.410702

Lessons Learned and Future Work

- Data Analysis is very hard
- Web scraping and beautiful soup
- Data can be hard to work with and I understand now the importance of data preprocessing
- I would like to collect more data for Utah and collect for other states to predict which areas will have an increase or decrease in housing sale prices
- It would also be interesting to look at data for housing all over the nation