Life is a Struggle

Group Members:

Yi Chen Huang(ID: yichenhu)

Cynthia Long(ID: meihual)

Emma Bu(ID: ybu)

Lily Yuan(ID: yuanzha2)

Content

Documentation

Our documentations are included in the two .py files:

- scraping.py: includes the scraping codes for the two websites: Apartments.com and Craigslist. It also includes the code of concatenating two excel files getting from two websites, merging them into the CSV file “houses.csv” to use in the menu.py file.
- menu.py: includes the codes of the user searching system, analysis results and the visualized plots. It used the CSV file “houses.csv” getting from scraping.py file.

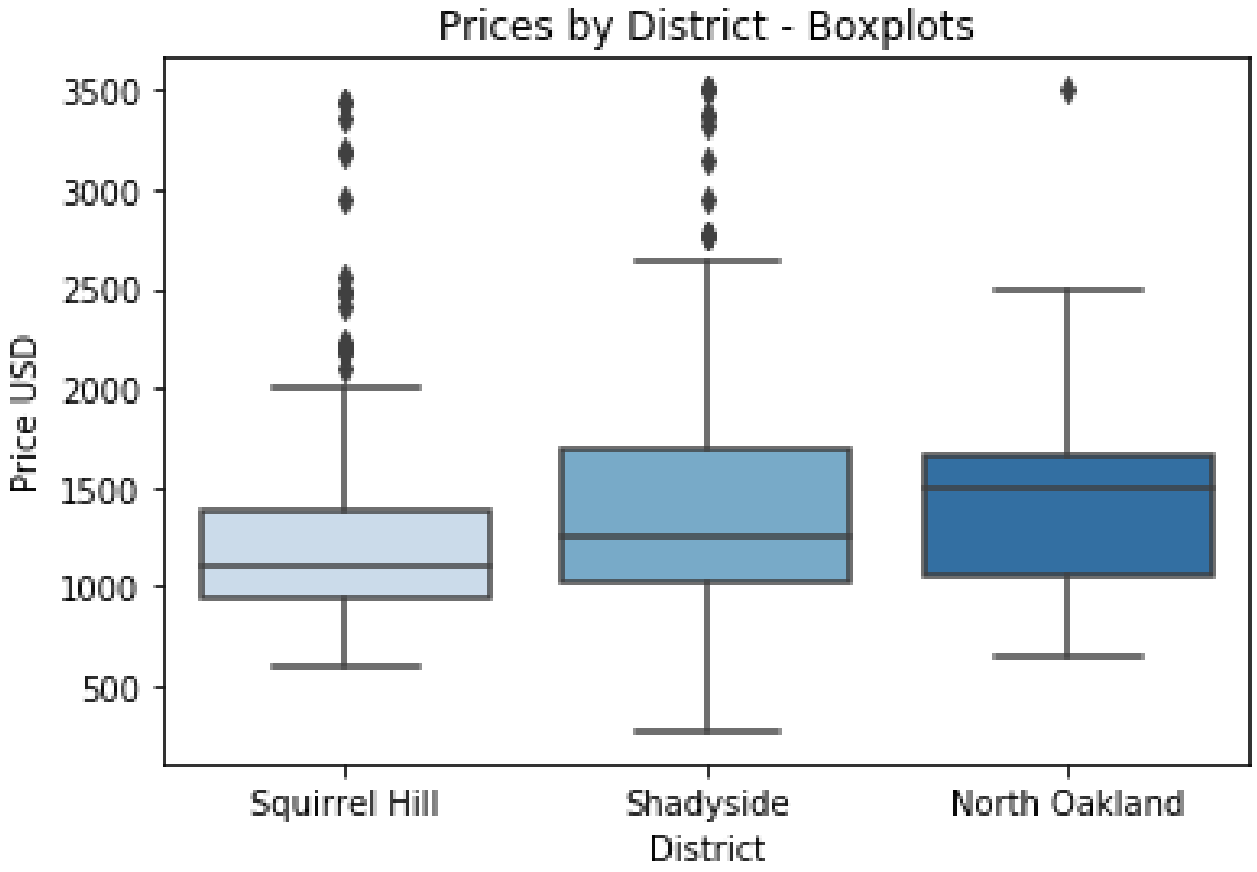
User Instruction

We designed a user searching system for the users to find their ideal house/apartment.

First, our user searching system would ask if the users want to show the statistical information on the housing in Pittsburgh.

If the user enters ‘y’, the system would display the descriptive price table of the three neighborhoods: Squirrel hill, North Oakland, Shadyside . And then show a boxplot that makes the users know clearly about the difference in housing prices of three neighborhood.

```
Do you want some statistical information on the housing in Pittsburgh?(y/n): y
               Counts         Average        Highest        Lowest
=============================================================================
Squirrel Hill  497            878.64         1779.50        300.00
North Oakland  99             702.67         1300.00        339.00
Shadyside      499            1036.95        2444.00        137.50
Overall        1095           934.87         2444.00        137.50
```


Prices by District - Boxplots
Price USD
3500
3000
2500
2000
1500
1000
500
Squirrel Hill
Shadyside
North Oakland
District

Next, showing the regression result that could guide the users to know which factors would influence the price the most. Thus, they could decide the factors that would influence the price of their ideal house like how many beds or bathrooms they would like to have, the size of the house etc.

```
Do you want more details on the effects of the factors?(y/n): y
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Price   R-squared:                       0.268
Model:                            OLS   Adj. R-squared:                  0.262
Method:                 Least Squares   F-statistic:                     40.07
Date:                Mon, 09 Dec 2019   Prob (F-statistic):           1.50e-67
Time:                        18:25:09   Log-Likelihood:                -12605.
No. Observations:                1103   AIC:                         2.523e+04
Df Residuals:                    1092   BIC:                         2.529e+04
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -4405.6927   2057.554     -2.141      0.032   -8442.900    -368.486
Beds       -6427.2640    918.842     -6.995      0.000   -8230.160   -4624.368
Baths       4844.2266   2103.532      2.303      0.021     716.805    8971.648
Size          15.9951      1.060     15.096      0.000      13.916      18.074
Distance    -455.8332     61.401     -7.424      0.000    -576.310    -335.356
NorthOakland -435.0589   2085.103     -0.209      0.835   -4526.321    3656.203
Shadyside   -775.0989   1133.619     -0.684      0.494   -2999.416    1449.218
SquirrelHill -3195.5349  1098.148     -2.910      0.004   -5350.255   -1040.815
Apartment   -454.5564   1546.214     -0.294      0.769   -3488.443    2579.331
Condo       -1.05e+04    1.6e+04     -0.657      0.511   -4.19e+04    2.09e+04
House        3.691e+04   4158.221      8.876      0.000    2.88e+04    4.51e+04
Townhomes   -4190.5436   8090.432     -0.518      0.605   -2.01e+04    1.17e+04
==============================================================================
Omnibus:                     2043.879   Durbin-Watson:                   1.661
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          4367436.408
Skew:                          12.732   Prob(JB):                         0.00
Kurtosis:                     310.216   Cond. No.                     1.16e+19
==============================================================================
```

Second, our system would ask if the users want more recommendations, if the users enter ‘y’, it would display some features that users could choose to find their ideal house/apartment.

The features users could select:

1. Neighborhood: Shadyside, North Oakland, Squirrel Hill or All areas(no preference)
2. Housing Type: house, apartment, condo, townhome, all the types(no preference)
3. Acceptable minimum housing price: use enter a numeric value
4. Affordable maximum housing price: use enter a numeric value
5. How to sort the results: price(high to low), price(low to high) and distance to CMU(in miles)
6. ALLDONE

```
Start a new search! Enter ALLDONE to quit.

Select the neighborhood:
1. Shadyside
2. North Oakland
3. Squirrel Hill
4. All the three areas

Please enter your choice: 1

Select the housing type:
1. House
2. Apartments
3. Condos
4. Townhomes
5. All the above types

Please enter your choice: 1
```

```
Enter the your acceptable minimum housing price: 0

Enter the your affordable maximum housing price: 9999
How do you want to sort the results
1. Price (high to low)
2. Price (low to high)
3. Distance to CMU (in miles)

Please enter your choice: 1
```

The result would show like below:

```
Find 45 places in total.
Here are the top 5 place for you
                                Address   District  Beds  Baths   Price  Size(sqft)                                                                                Website  Distance(miles)   Type
0                                 Aiken  Shadyside   3.0    1.5  2300.0      1700.0  https://pittsburgh.craigslist.org/apa/d/pittsburgh-3br-house-for-rent-close-to/7029194632.html             1.3  House
1          234 Albert St near Boggs Ave.  Shadyside   2.0    1.0  1100.0      1200.0      https://pittsburgh.craigslist.org/apa/d/pittsburgh-spacious-2br-home-in-mt/7015640569.html             7.3  House
2                       3232 Parkview Ave  Shadyside   4.0    2.0  1800.0      1250.0  https://pittsburgh.craigslist.org/apa/d/pittsburgh-beautiful-4-br-2-bath-house/7028961521.html             1.1  House
3  6426 Howe Street near Festival Street  Shadyside   2.0    2.5  2600.0      1982.0  https://pittsburgh.craigslist.org/apa/d/pittsburgh-quaint-cottage-from-january/7027490530.html             2.8  House
4                    crescent hills road  Shadyside   3.0    1.5  1795.0      1550.0   https://pittsburgh.craigslist.org/apa/d/pittsburgh-good-credit-income-rent-to/7016494828.html            13.7  House
```

The first line shows the total available items we found in to websites. And then we provide the top five places based on the users' selection and their preference of sorting the results. Moreover, the system would display a scatter plot based on the searching result and output a map.html based on the top five places.

In the end, the system would ask if the users want to show all the searched item. If the user type 'y', the system would display all the housing items' details we found(not only the top five).

The menu would continue displaying until the users enter ALLDONE.

Abstract

- Utilize BeautifulSoup in Python to scrap 2000+ housing data such as location, price, house type from Craigslist and Apartment.com
- Manipulate re package to clean the raw data into pandas Dataframe and concatenate the data of two websites
- Design interactive search program to assist target users excavate ideal housing options based on price and distance
- Visualize statistics data by using matplotlib to draw box plot, scatter plot, regression plot and map