Carnegie Mellon University

Pittsburgh Bike Data Analysis

JUN 01, 2023

Hsiu-Yuan Yang

Agenda

1. Background & Objective
2. Data Source & Data Preparation
3. EDA & Visualization
4. Model Analysis

Background & Objective

Background & Objective

Background: The Rise of Bike Transportation

The popularity of shared mobility services has led Bike Share Pittsburgh Inc (POGOH) to launch a bike sharing program in Pittsburgh, Pennsylvania. The program aims to provide residents and visitors with accessible, sustainable, and affordable bicycle rentals, integrating human-powered transportation as an essential component of the larger public transit system. By doing so, the program offers greater convenience and transportation opportunities for Pittsburgh residents while also providing an additional sightseeing option for visitors.

However, the distribution of bike stations has become a subject of debate, with concerns about some neighborhoods being underserved and others having an excess of stations.

Background & Objective (Cont'd)

Objective: To Access the Current Design of Pittsburgh Bike Stations

In this project, we aim to examine the relationship between the underlying census data and the number of bicycle rental stations in different census tracts in Pittsburgh, with a focus on factors such as population, race, income, education level, transportation characteristics, etc.

The purpose of this work is to reassess the current design of bike stations in Pittsburgh and determine if improvements or adjustments are necessary. The primary goal is to identify any disparities and propose changes to enhance the bike sharing program, making it more equitable and efficient.

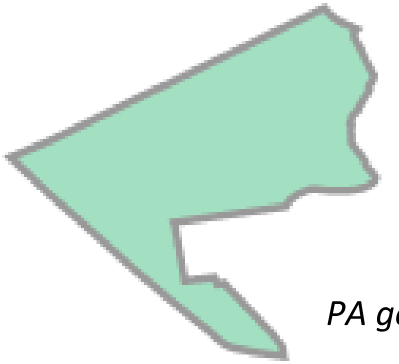
Data Source

Data Source

1. Pittsburgh Geographical Data

TIGER/Line Shapefiles from US Census Bureau (2020)[1]

Pennsylvania shapefiles was used as the base and extraction / processing was done to filter out Pittsburgh shapefiles



PA geometry shape

2. Pittsburgh Census Data

Population details / info from US Census Bureau (2020)[2]

On top of extraction of the census data, preprocessing was also performed for easier analysis purpose, including the following:

1. Population density was calculated
2. A new category was created (i.e. college_degree_or_above) to simplify the education level grouping

Notes:

1. *For more information, please see https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.2020.html#list-tab-790442341.*
2. *For more information, please see https://data.census.gov/.*

Data Source (Cont'd)

3. Pittsburgh Bike Station Data[1]

Bike station data provided from Healthy Ride (2021)[2]

A snapshot of station locations and capacities

4. Pittsburgh Bike Station Activity Data[1]

Bike rental data provided from Healthy Ride (2021)[2]

Details of the rental trips, where a trip is defined as any rental longer than a minute that begins and ends at a valid Healthy Ride station

Healthy Ride

powered by nextbike

Notes:

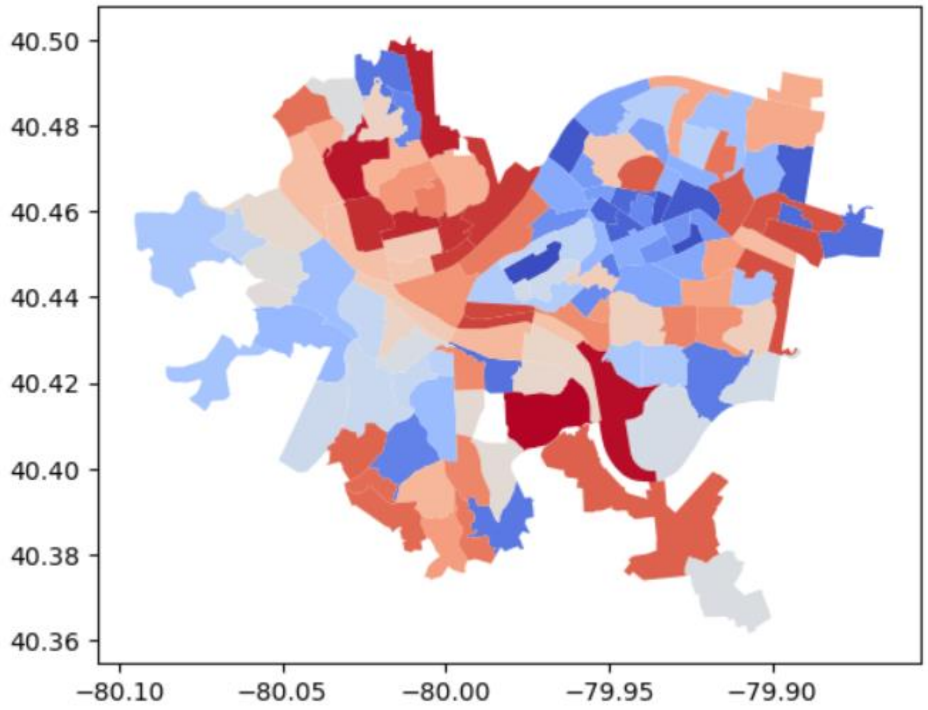
1. *For more information, please see https://healthyridepgh.com/data/ .*
2. *While Healthy Ride has stored their data by quarterly basis, for this study, we have consolidated the data to account for the whole 2021.*

Data Source (Cont'd)

1. Pittsburgh Geographical Data[1]

The geographical data contains the following information[2]:

- State Federal Information Processing System (FIPS) code
- Country FIPS code
- Census Tract Code
- GEOID (serves as census block identifiers)
- Census Land Area
- Census Water Area
- Census latitude of the internal point
- Census longitude of the internal point
- Geometry representation



Pittsburgh geometry map

Notes:
1. *Our data scope is for year 2020.*
2. *The list is not exhaustive; only information that are used in this project are listed.*

Data Source (Cont'd)

2. Pittsburgh Census Data

The census data used can be divided into the below categories:

General Overview	Transportation Related
Examples: • Population density • Gender • Median income • Median house value • Education level • Race	Examples: • Public transportation for work • Travel time to work

Data Source (Cont'd)

2. Pittsburgh Census Data (Cont'd)

The full set of census data attributes (after preprocessing) used are as follows:

```
1   NAME
2   POPULATION
3   NATIVE
4   AGE_UNDER_17
5   AGE_18_TO_64
6   AGE_OVER_65
7   MALE
8   FEMALE
9   ONE_RACE
10  TWO_RACES
11  WHITE
12  BLACK_AFRICAN_AMERICAN
13  AMERICAN_INDIAN_ALASKA_NATIVE
14  ASIAN
15  HAWAIIAN_PACIFIC
16  OTHER_RACE
17  HISPANIC_LATINO
18  LANGUAGE_OTHER_THAN_ENGLISH
19  NEVER_MARRIED
20  NOW_MARRIED
21  DIVORCED_SEPARATED
22  WIDOWED
23  LESS_THAN_HIGH_SCHOOL
24  HIGH_SCHOOL_GRADUATE
25  COLLEGE_ASSOCIATE_DEGREE
26  BACHELOR_DEGREE
27  GRADUATE_PROFESSIONAL_DEGREE
28  INDIVIDUAL_INCOME_MEDIAN
29  MOVED_WITHIN_SAME_COUNTY
30  MOVED_FROM_DIFFERENT_COUNTY
31  MOVED_FROM_DIFFERENT_STATE
32  MOVED_FROM_ABROAD
33  WORKER
34  WORKER_NOT_WFH
35  WORKER_DEPART_0000_0559
36  WORKER_DEPART_0600_0729
37  WORKER_DEPART_0730_0859
38  WORKER_DEPART_0900_2359
39  TRAVEL_TIME_TO_WORK_LESS_THAN_10_MIN
40  TRAVEL_TIME_TO_WORK_10_TO_14_MIN
41  TRAVEL_TIME_TO_WORK_15_TO_19_MIN
42  TRAVEL_TIME_TO_WORK_20_TO_24_MIN
43  TRAVEL_TIME_TO_WORK_25_TO_29_MIN
44  TRAVEL_TIME_TO_WORK_30_TO_34_MIN
45  TRAVEL_TIME_TO_WORK_35_TO_44_MIN
46  TRAVEL_TIME_TO_WORK_45_TO_59_MIN
47  TRAVEL_TIME_TO_WORK_MORE_THAN_60_MIN
48  WORKER_VEHICLE_DROVE_ALONE
49  WORKER_VEHICLE_CARPOOL
50  WORKER_PUBLIC_TRANSPORTATION
51  HOUSEHOLD
52  HOUSEHOLD_SIZE
53  FAMILY
54  FAMILY_SIZE
55  HOUSEHOLD_WITH_CHILREN
56  HOUSEHOLD_MARRIED_COUPLE
57  HOUSEHOLD_MALE_NO_SPOUSE
58  HOUSEHOLD_FEMALE_NO_SPOUSE
59  HOUSEHOLD_NONFAMILY
60  POVERTY
61  DISABILITY
62  MEDIAN_HOUSE_VALUE
63  Area
64  POPULATION_DENSITY
65  COLLEGE_DEGREE_OR_ABOVE
```

Data Source (Cont'd)

3. Pittsburgh Bike Station Data[1]

The bike station data contains the following information:

- Station ID
- Station name
- Latitude / longitude coordinates
- Number of individual docking points at each station

During pre-processing, it was noted that the station data for 2021 Q2 was corrupted and the information stored are not aligned with other station data.

Upon examining the rental data in 2021 Q2, it was determined that the station data for 2021 Q2 is more aligned with the station data for 2021 Q3 instead of 2021 Q1, hence, we have used the Q3 file as a substitute.

It is noted that there are out of the 128 census tracts, only 42 of them have bike stations.

Notes:

1. *Our data scope is for year 2021.*

Data Source (Cont'd)

4. Pittsburgh Bike Station Activity Data

The bike station activity (rental) data contains the following information:

- Trip ID
- Bike ID
- Trip start day ad time
- Trip end day and time
- Trip duration (in seconds)
- Trip start station name and station ID[1]
- Trip end station name and station ID[1]
- User type

For easier interpretation purpose, we decided to look at bike activity from demand and supply angle.

Therefore, we have rendered the trip start station name and station ID as "demand" and the trip end station name and station ID as "supply".

Demand counts and supply counts are then be aggregated to be utilized in the following analysis.

Notes:
1. *These are the fields that have been use in this study.*

Data Preparation

Data Preparation

Consolidation relevant information

To facilitate the regression / modelling, the below actions were executed:

1. Bike station data and bike station activity data were consolidated with census data and geographical data

2. While bike station activity data is in quarterly basis, the values have been aggregated to account for the total activity in 2021

Data Preparation (Cont'd)

Final Dataset Used for Modeling

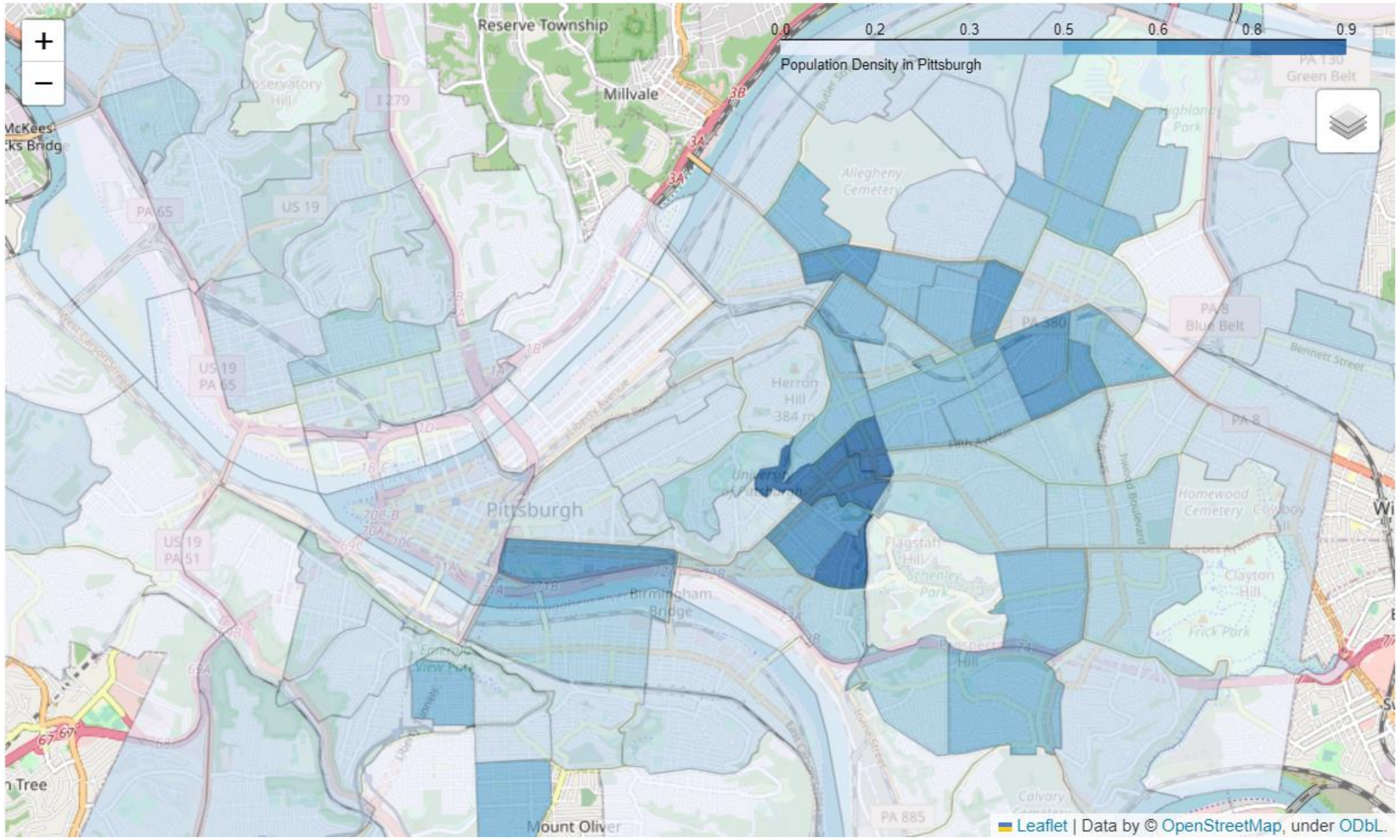
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 128 entries, 0 to 127
Data columns (total 74 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   GEO_ID                              128 non-null    object
 1   NAME                                128 non-null    object
 2   POPULATION                          128 non-null    int64
 3   NATIVE                              128 non-null    float64
 4   AGE_UNDER_17                        128 non-null    float64
 5   AGE_18_TO_64                        128 non-null    float64
 6   AGE_OVER_65                         128 non-null    float64
 7   MALE                                128 non-null    float64
 8   FEMALE                              128 non-null    float64
 9   ONE_RACE                            128 non-null    float64
 10  TWO_RACES                           128 non-null    float64
 11  WHITE                               128 non-null    float64
 12  BLACK_AFRICAN_AMERICAN              128 non-null    float64
 13  AMERICAN_INDIAN_ALASKA_NATIVE       128 non-null    float64
 14  ASIAN                               128 non-null    float64
 15  HAWAIIAN_PACIFIC                    128 non-null    float64
 16  OTHER_RACE                          128 non-null    float64
 17  HISPANIC_LATINO                     128 non-null    float64
 18  LANGUAGE_OTHER_THAN_ENGLISH         128 non-null    float64
 19  NEVER_MARRIED                       128 non-null    float64
 20  NOW_MARRIED                         128 non-null    float64
 21  DIVORCED_SEPARATED                  128 non-null    float64
 22  WIDOWED                             128 non-null    float64
 23  LESS_THAN_HIGH_SCHOOL               128 non-null    float64
 24  HIGH_SCHOOL_GRADUATE                128 non-null    float64
 25  COLLEGE_ASSOCIATE_DEGREE            128 non-null    float64
 26  BACHELOR_DEGREE                     128 non-null    float64
 27  GRADUATE_PROFESSIONAL_DEGREE        128 non-null    float64
 28  INDIVIDUAL_INCOME_MEDIAN            128 non-null    int64
 29  MOVED_WITHIN_SAME_COUNTY            128 non-null    float64
 30  MOVED_FROM_DIFFERENT_COUNTY         128 non-null    float64
 31  MOVED_FROM_DIFFERENT_STATE          128 non-null    float64
 32  MOVED_FROM_ABROAD                   128 non-null    float64
 33  WORKER                              128 non-null    int64
 34  WORKER_NOT_WFH                      128 non-null    int64
 35  WORKER_DEPART_0000_0559             128 non-null    float64
 36  WORKER_DEPART_0600_0729             128 non-null    float64
 37  WORKER_DEPART_0730_0859             128 non-null    float64
 38  WORKER_DEPART_0900_2359             128 non-null    float64
 39  TRAVEL_TIME_TO_WORK_LESS_THAN_10_MIN  128 non-null  float64
 40  TRAVEL_TIME_TO_WORK_10_TO_14_MIN    128 non-null    float64
 41  TRAVEL_TIME_TO_WORK_15_TO_19_MIN    128 non-null    float64
 42  TRAVEL_TIME_TO_WORK_20_TO_24_MIN    128 non-null    float64
 43  TRAVEL_TIME_TO_WORK_25_TO_29_MIN    128 non-null    float64
 44  TRAVEL_TIME_TO_WORK_30_TO_34_MIN    128 non-null    float64
 45  TRAVEL_TIME_TO_WORK_35_TO_44_MIN    128 non-null    float64
 46  TRAVEL_TIME_TO_WORK_45_TO_59_MIN    128 non-null    float64
 47  TRAVEL_TIME_TO_WORK_MORE_THAN_60_MIN  128 non-null  float64
 48  WORKER_VEHICLE_DROVE_ALONE          128 non-null    int64
 49  WORKER_VEHICLE_CARPOOL              128 non-null    int64
 50  WORKER_PUBLIC_TRANSPORTATION        128 non-null    int64
 51  HOUSEHOLD                           128 non-null    int64
 52  HOUSEHOLD_SIZE                      128 non-null    float64
 53  FAMILY                              128 non-null    int64
 54  FAMILY_SIZE                         128 non-null    float64
 55  HOUSEHOLD_WITH_CHILREN              128 non-null    int64
 56  HOUSEHOLD_MARRIED_COUPLE            128 non-null    int64
 57  HOUSEHOLD_MALE_NO_SPOUSE            128 non-null    int64
 58  HOUSEHOLD_FEMALE_NO_SPOUSE          128 non-null    int64
 59  HOUSEHOLD_NONFAMILY                 128 non-null    int64
 60  POVERTY                             128 non-null    int64
 61  DISABILITY                          128 non-null    int64
 62  MEDIAN_HOUSE_VALUE                  128 non-null    int64
 63  Area                                128 non-null    int64
 64  POPULATION_DENSITY                  128 non-null    float64
 65  COLLEGE_DEGREE_OR_ABOVE             128 non-null    float64
 66  bike_station_count                  128 non-null    int64
 67  INTPTLAT                            128 non-null    float64
 68  INTPTLON                            128 non-null    float64
 69  activity_count_Q1                   128 non-null    float64
 70  activity_count_Q2                   128 non-null    float64
 71  activity_count_Q3                   128 non-null    float64
 72  activity_count_Q4                   128 non-null    float64
 73  total_activity_count                128 non-null    float64
dtypes: float64(53), int64(19), object(2)
memory usage: 74.1+ KB
```

EDA & Visualization

Carnegie Mellon University

Pittsburgh Census Characteristics

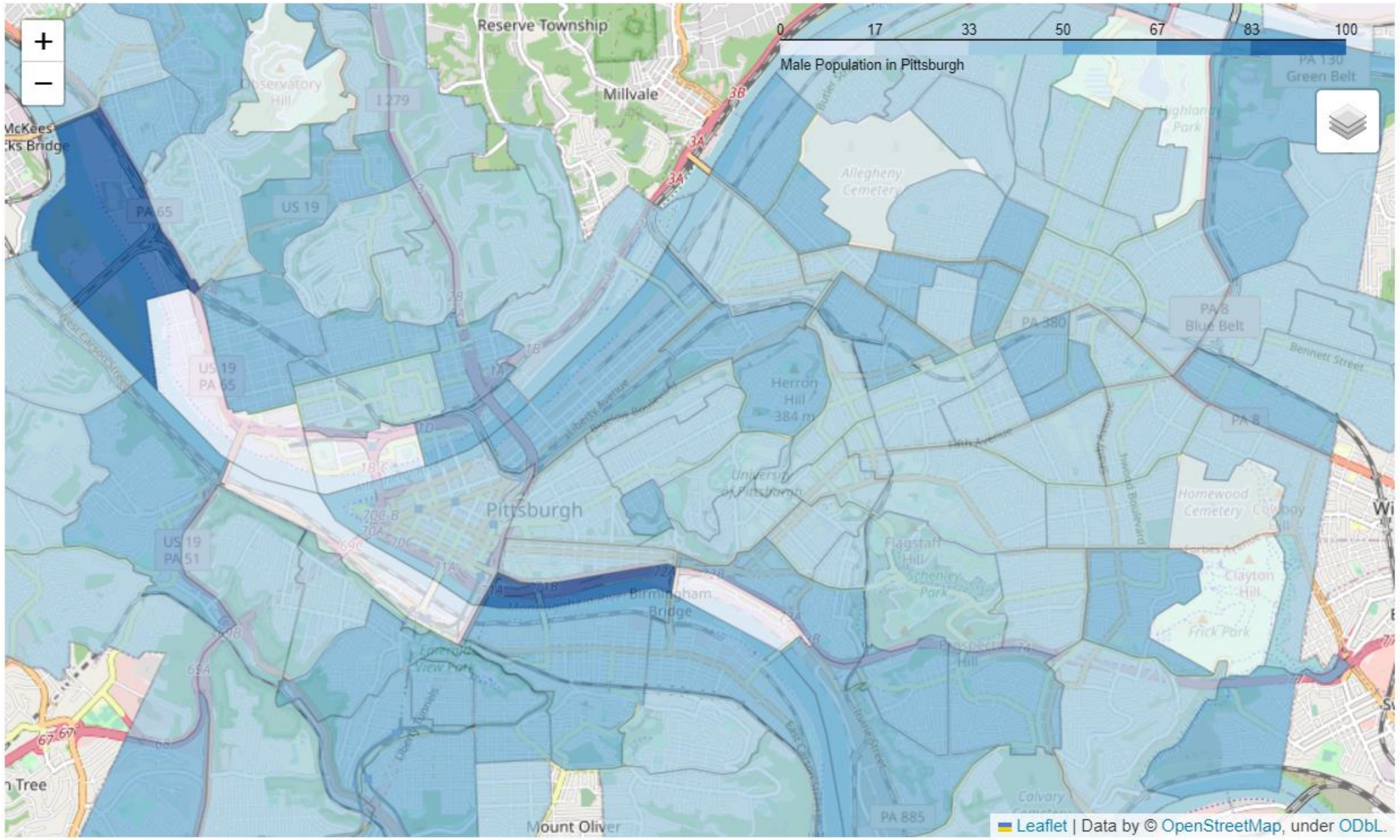
Population Density



Carnegie Mellon University

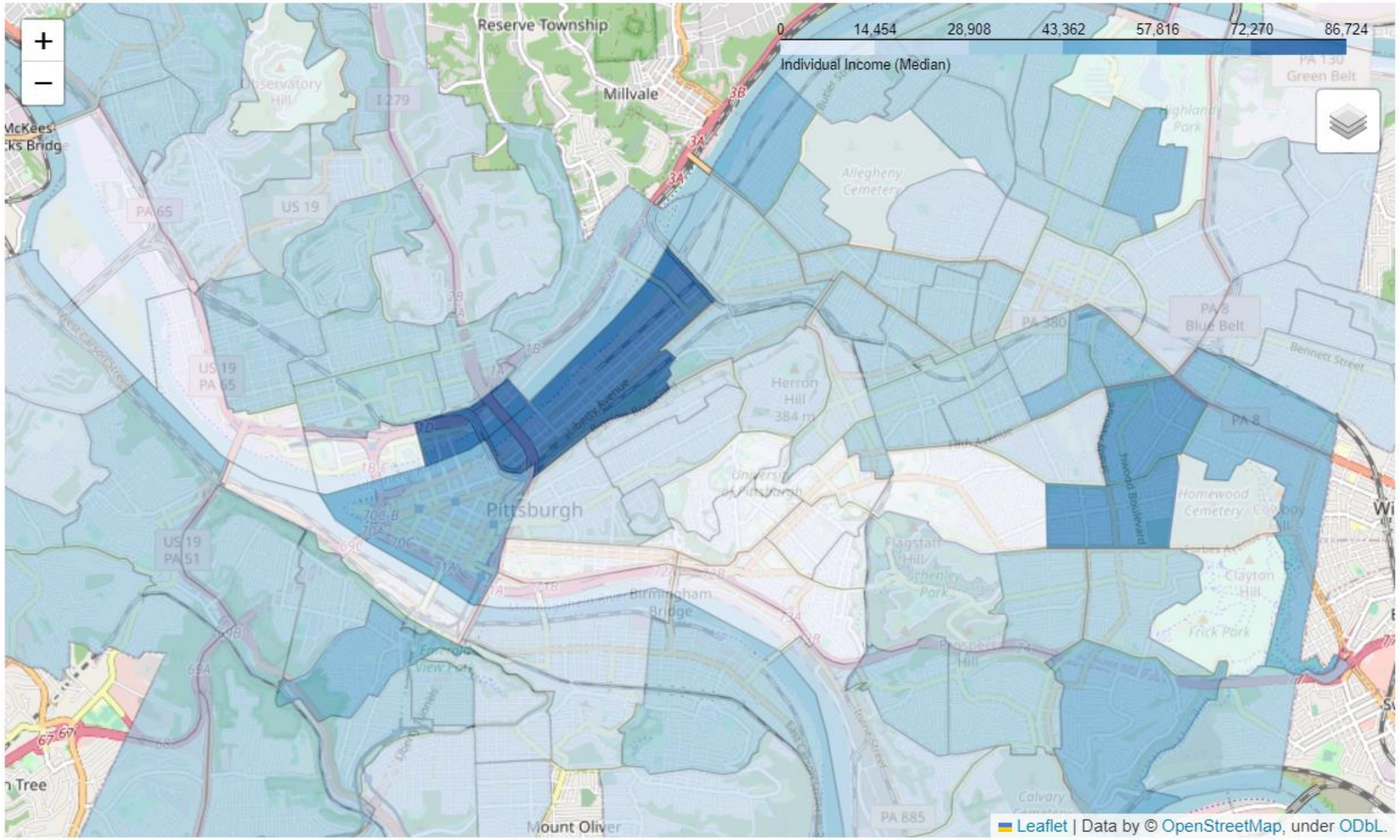
Pittsburgh Census Characteristics (Cont'd)

Gender



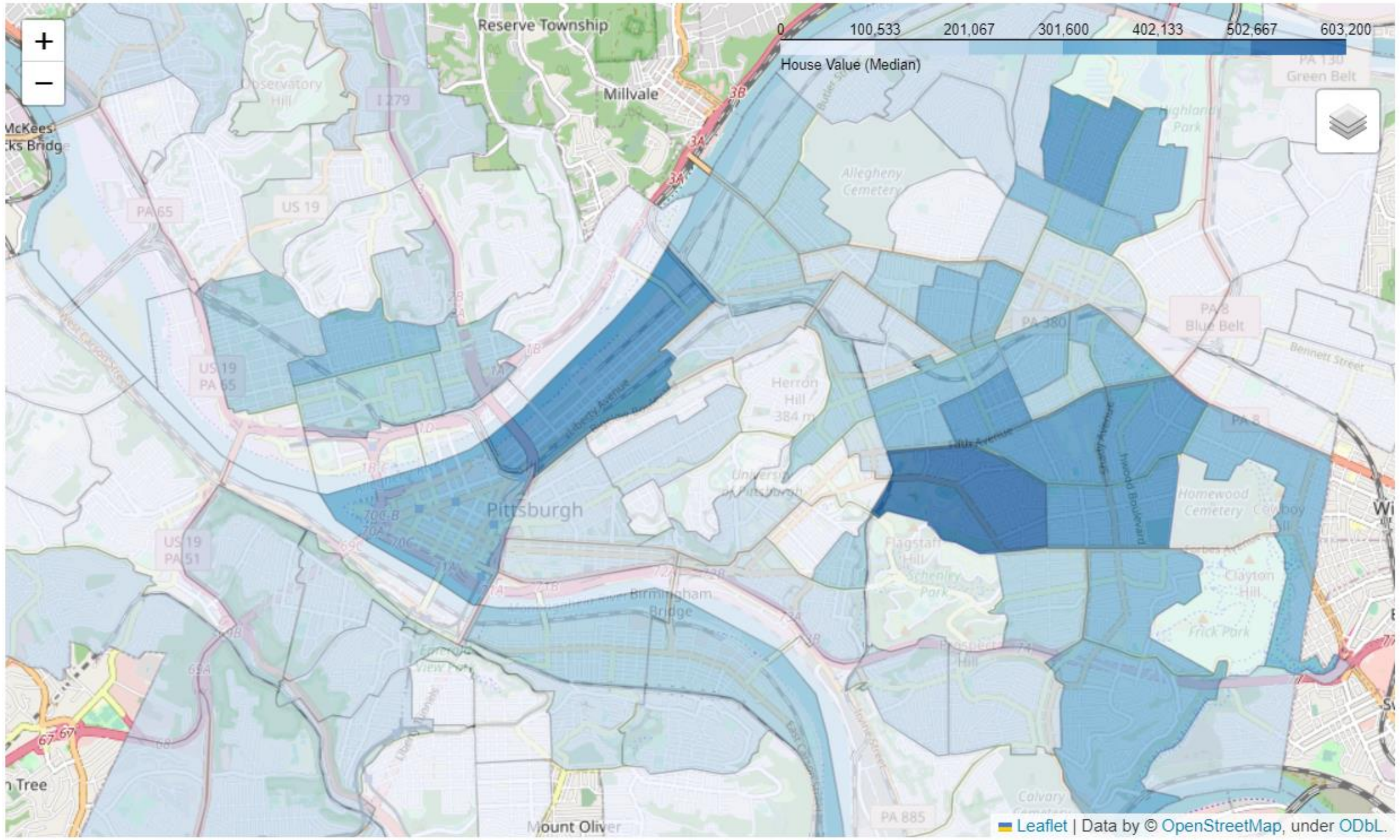
Pittsburgh Census Characteristics (Cont'd)

Individual Median Income



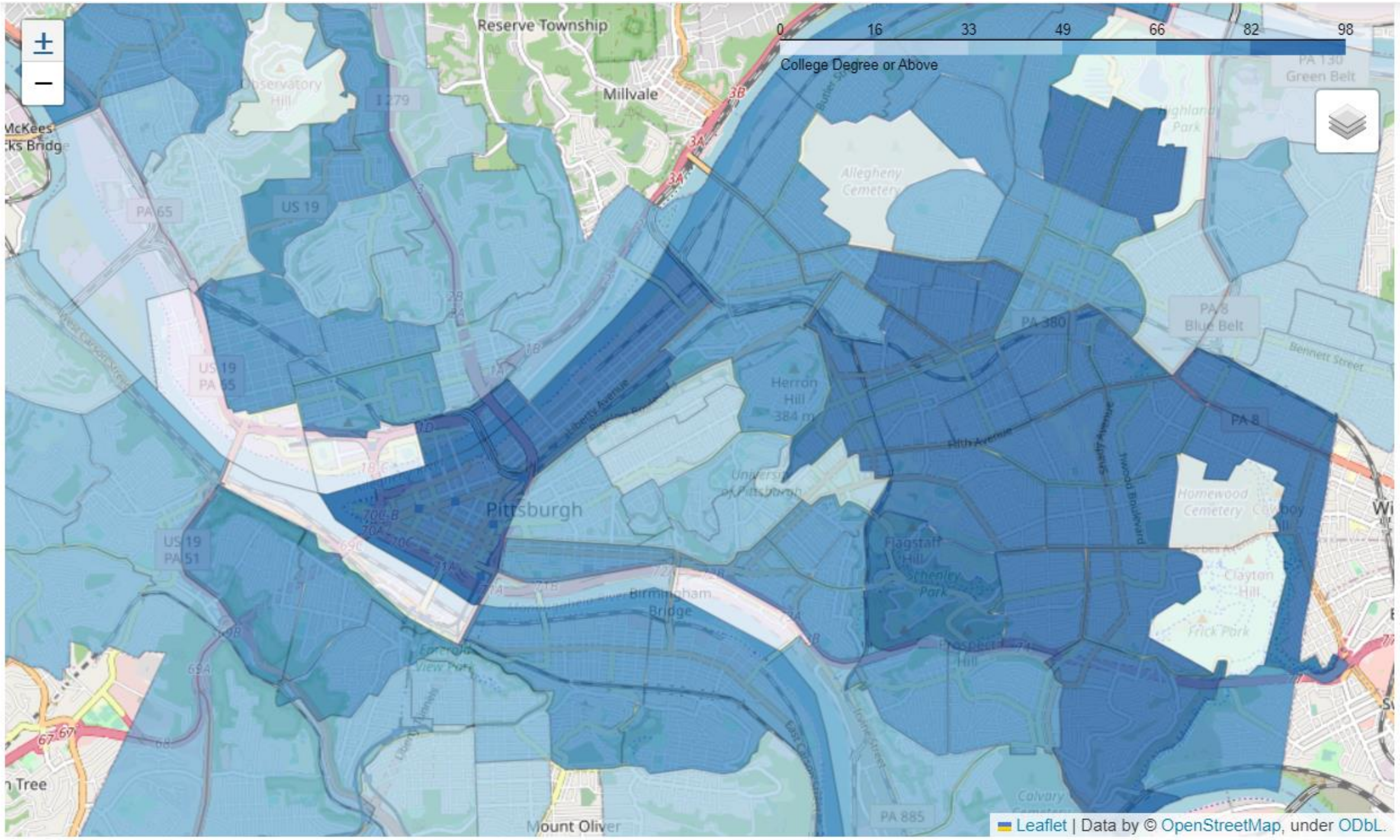
Pittsburgh Census Characteristics (Cont'd)

Median House Value



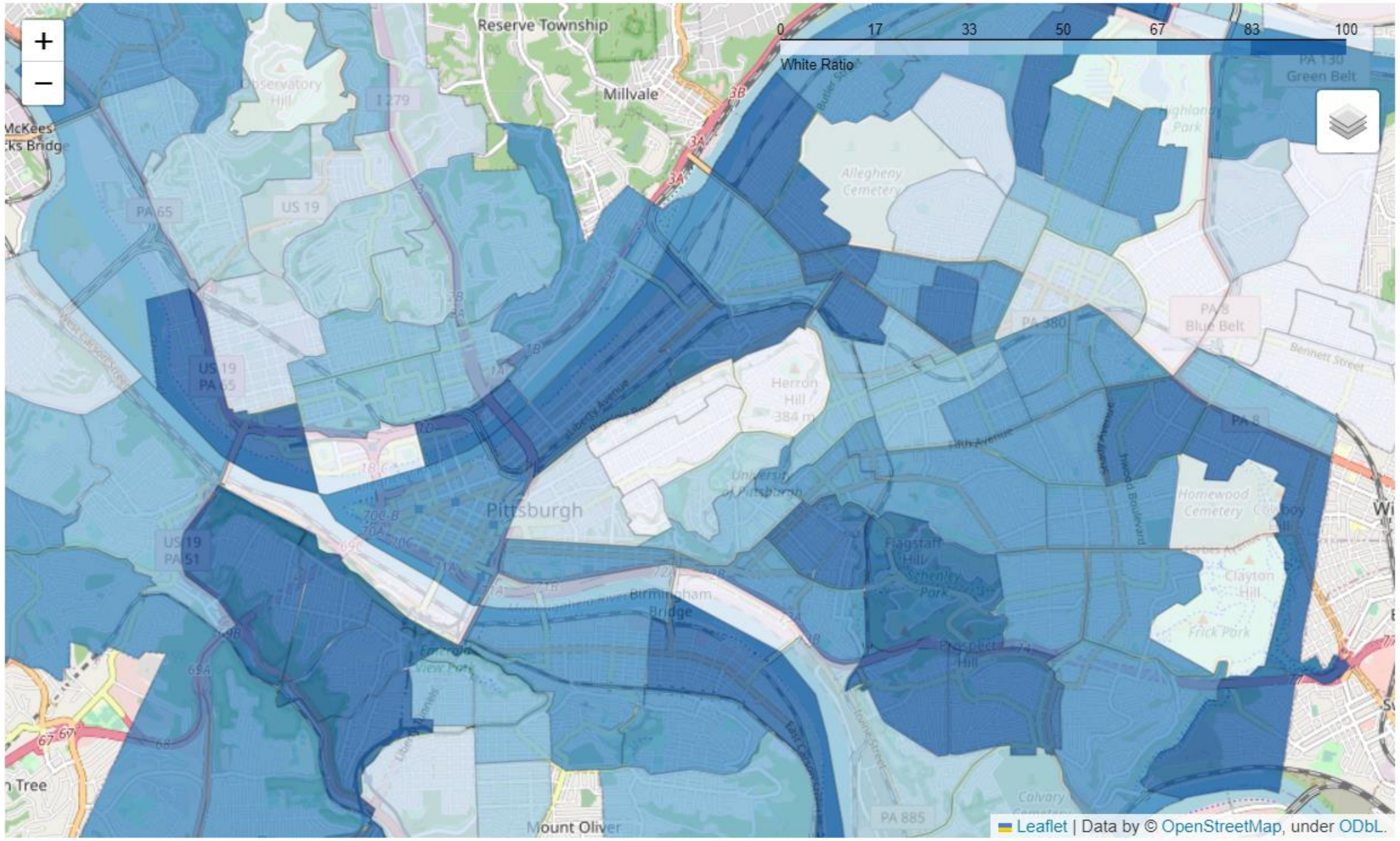
Pittsburgh Census Characteristics (Cont'd)

Education Level (College Degree or Above)



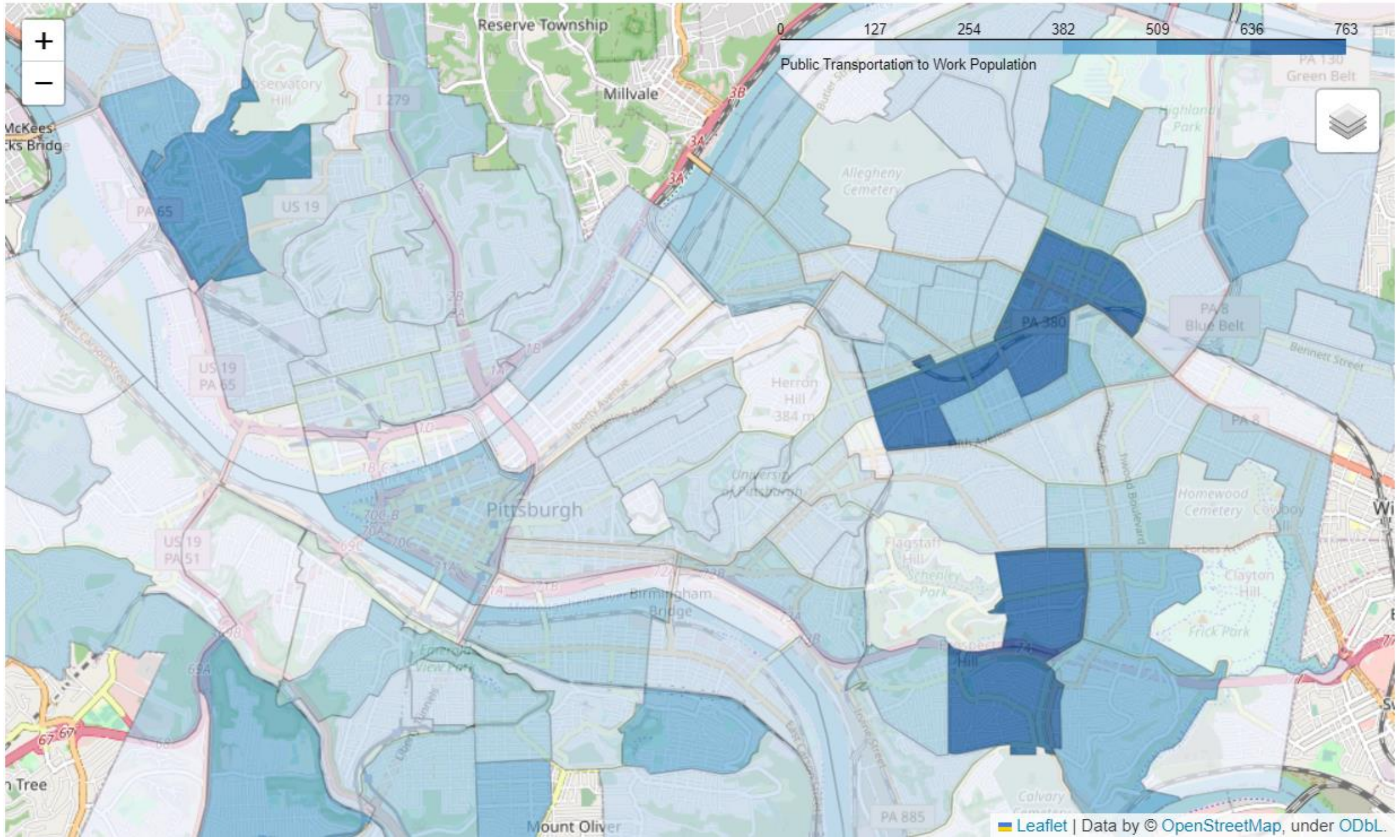
Pittsburgh Census Characteristics (Cont'd)

Race (White Ratio)



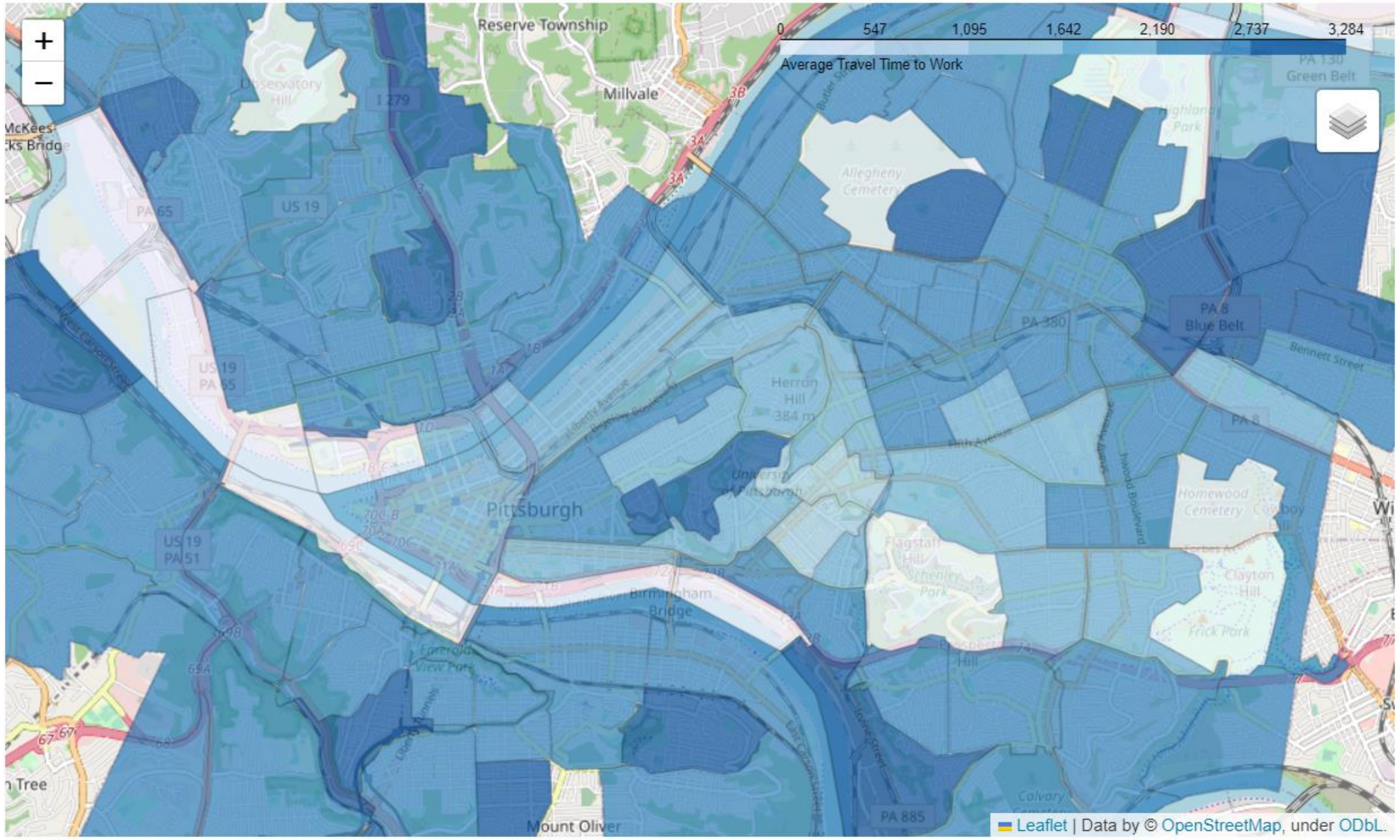
Pittsburgh Census Characteristics (Cont'd)

Usage of Public Transportation to Work Ratio

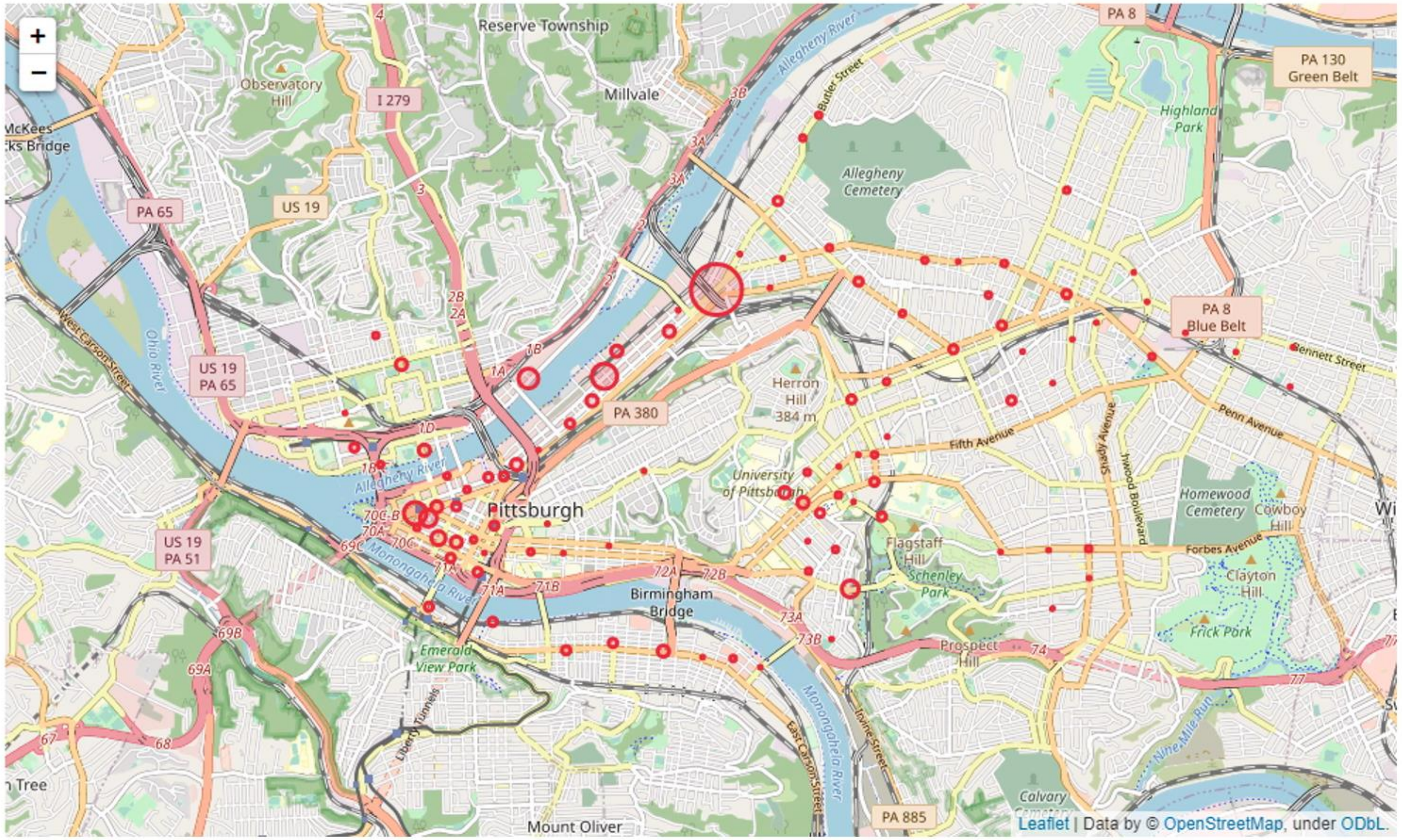


Pittsburgh Census Characteristics (Cont’d)

Average Travel Time to Work (mins)

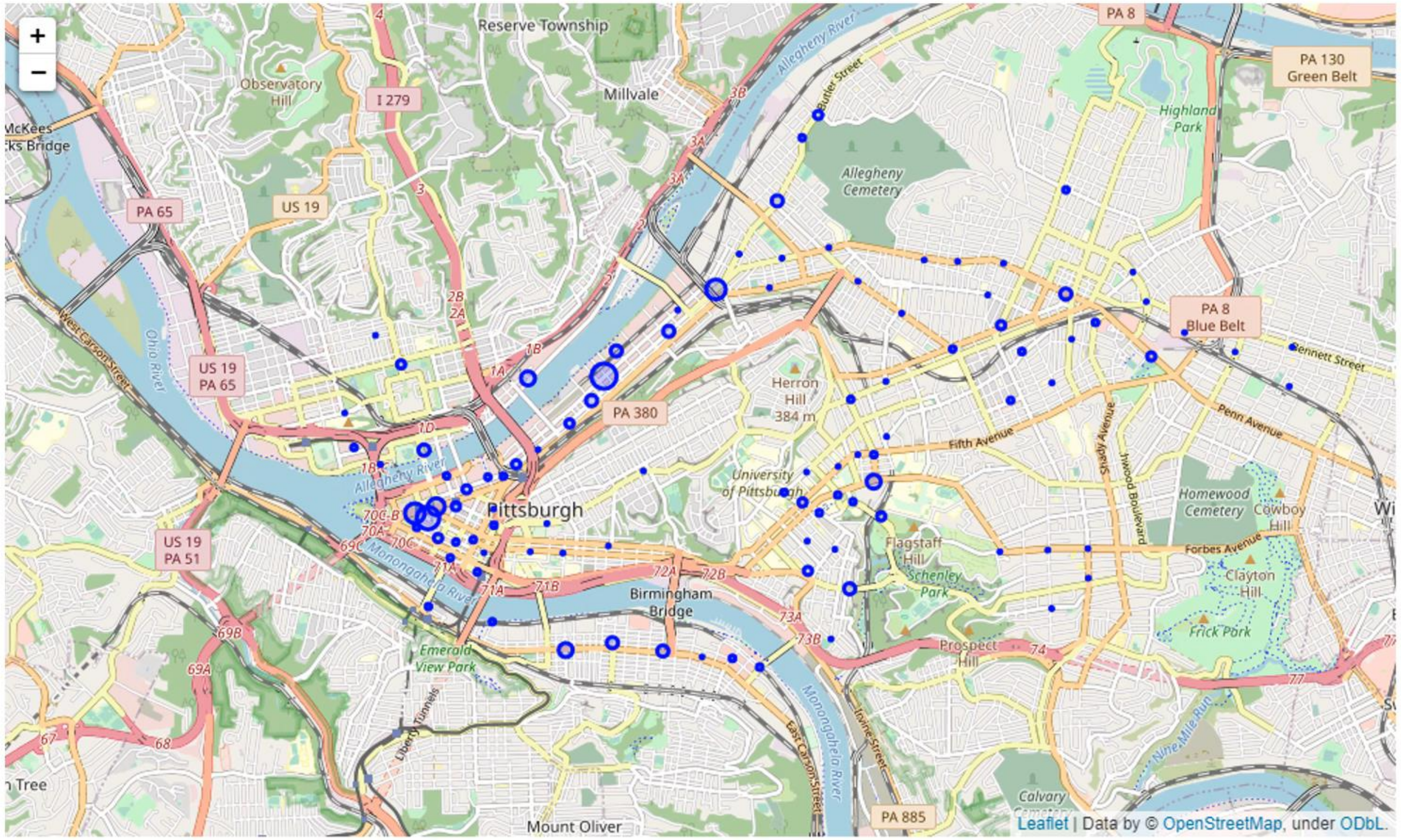


Demand (Bike Activities) in 2020



Demand Map for January 2020

Supply (Bike Activities) in 2020



Supply Map for January 2020

Carnegie Mellon University

Model Analysis

Our Models

To Examine the Current Design of Bike Stations

Below are the 2 models defined in this analysis[1]:

1. "Number of bike stations" versus "census features"
2. "Average activity[2]" versus "number of bike stations & census features"

2 types of methods are utilized:

1. Spatial regression
2. Elastic net[3]

Notes:
1. *The data unit is per census tract.*
2. *Activity is defined as the "demand" count plus "supply" count in each bike station. Average activity refers to the total activity count for all bike stations divided by the total number of bike station counts within one census tract.*
3. *Considering that we only have a limited sample size (only 42 census tracts have bike stations) and a large number of census features, using spatial regression did not give us too meaningful insights, hence, elastic net was also used for elimination of features.*

Spatial Regression

1. "Number of bike stations" versus "census features"

(a) Regression using the full set of features:

```
# full features first
# get X and y (for full set)
feature_cols = ['POPULATION_DENSITY', "AGE_UNDER_17", "AGE_18_TO_64", "AGE_OVER_65", "MALE", 'WHITE',
                "BLACK_AFRICAN_AMERICAN", "AMERICAN_INDIAN_ALASKA_NATIVE", "ASIAN", "HAWAIIAN_PACIFIC",
                "HISPANIC_LATINO", "LANGUAGE_OTHER_THAN_ENGLISH", 'MEDIAN_HOUSE_VALUE',
                'COLLEGE_DEGREE_OR_ABOVE', "INDIVIDUAL_INCOME_MEDIAN", "WORKER_NOT_WFH",
                "WORKER_DEPART_0000_0559", "WORKER_DEPART_0600_0729", "WORKER_DEPART_0730_0859", "WORKER_DEPART_0900_2359",
                "TRAVEL_TIME_TO_WORK_LESS_THAN_10_MIN", 'TRAVEL_TIME_TO_WORK_10_TO_14_MIN',
                'TRAVEL_TIME_TO_WORK_15_TO_19_MIN', 'TRAVEL_TIME_TO_WORK_20_TO_24_MIN',
                'TRAVEL_TIME_TO_WORK_25_TO_29_MIN', 'TRAVEL_TIME_TO_WORK_30_TO_34_MIN',
                'TRAVEL_TIME_TO_WORK_35_TO_44_MIN', 'TRAVEL_TIME_TO_WORK_45_TO_59_MIN',
                'TRAVEL_TIME_TO_WORK_MORE_THAN_60_MIN', 'WORKER_VEHICLE_DROVE_ALONE',
                'WORKER_VEHICLE_CARPOOL', 'WORKER_PUBLIC_TRANSPORTATION', 'POVERTY', 'DISABILITY']
X = census_df[feature_cols].values

# standardization first
scaler = StandardScaler()
X = scaler.fit_transform(X)
y = census_df.bike_station_count.values

ols_spreg = spreg.OLS(y, X, w=W, spat_diag=True, moran = True)
print(ols_spreg.summary)
```

Python command

The spatial weights were calculated leveraging libpysal library with the below command:

```
# calculate weights
points = list(zip(census_df['INTPTLAT'], census_df['INTPTLON']))
W=libpysal.weights.DistanceBand(points,threshold=11.2,binary=False)
```

Spatial Regression (Cont'd)

1. "Number of bike stations" versus "census features" (Cont'd)

(a) Regression using the full set of features (Cont'd):

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  :     dep_var                Number of Observations:         128
Mean dependent var  :      0.7812                Number of Variables   :          35
S.D. dependent var  :      1.9398                Degrees of Freedom    :          93
R-squared           :      0.7514
Adjusted R-squared  :      0.6605
Sum squared residual:     118.809                F-statistic           :      8.2667
Sigma-square        :       1.278                Prob(F-statistic)     :   2.867e-16
S.E. of regression  :       1.130                Log likelihood        :    -176.855
Sigma-square ML     :       0.928                Akaike info criterion :     423.710
S.E of regression ML:      0.9634                Schwarz criterion     :     523.531

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT       0.7812500       0.0999027       7.8201073       0.0000000
               var_1      -0.1096981       0.1601751      -0.6848638       0.4951332
               var_2      -0.3921096       0.4277514      -0.9166765       0.3616832
               var_3      -0.4986081       0.9758415      -0.5109520       0.6105955
               var_4      -0.3862427       0.4488300      -0.8605546       0.3916965
               var_5      -0.1847496       0.1972263      -0.9367391       0.3513190
               var_6       0.8712260       1.2818038       0.6796875       0.4983907
               var_7       1.3133771       1.1152717       1.1776297       0.2419491
               var_8      -0.2977766       0.1185095      -2.5126814       0.0137037
               var_9       0.3727325       0.3473711       1.0730096       0.2860432
              var_10       0.0290473       0.1193684       0.2433411       0.8082771
              var_11       0.0117169       0.1525549       0.0768045       0.9389440
              var_12       0.0056149       0.2744463       0.0204589       0.9837211
              var_13      -0.0238102       0.2015485      -0.1181366       0.9062142
              var_14      -0.5881781       0.3203899      -1.8358197       0.0695798
              var_15       1.0693322       0.2321361       4.6064888       0.0000130
              var_16       4.9003185       0.6375255       7.6864662       0.0000000
              var_17      -4.5695019       6.3733229      -0.7169732       0.4751875
              var_18      -8.6417610      11.9703565      -0.7219301       0.4721487
              var_19     -10.8193133      14.2235522      -0.7606618       0.4487834
              var_20     -10.9445569      14.6108401      -0.7490710       0.4557049
              var_21       5.9521877       7.4428575       0.7997181       0.4259121
              var_22       5.2280661       6.9375451       0.7535902       0.4529990
              var_23       5.7203572       8.0773125       0.7082006       0.4805920
              var_24       6.0782112       8.3839729       0.7249798       0.4702846
              var_25       2.9990909       4.1548672       0.7218259       0.4722125
              var_26       5.8878780       8.0225248       0.7339183       0.4648448
              var_27       2.8343018       3.9086490       0.7251359       0.4701893
              var_28       2.3071508       3.2412609       0.7118066       0.4783663
              var_29       3.6458718       5.0970389       0.7152921       0.4762205
              var_30      -2.8545615       0.4041137      -7.0637584       0.0000000
              var_31      -0.5480127       0.1618232      -3.3864902       0.0010388
              var_32      -0.7193145       0.2363374      -3.0435916       0.0030394
              var_33      -0.9016902       0.4455309      -2.0238556       0.0458542
              var_34       0.0613769       0.1331994       0.4607895       0.6460254
------------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER          982.124

TEST ON NORMALITY OF ERRORS
TEST                             DF        VALUE           PROB
Jarque-Bera                       2        14.380          0.0008

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                             DF        VALUE           PROB
Breusch-Pagan test               34        102.443         0.0000
Koenker-Bassett test             34         68.455         0.0004

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                           MI/DF       VALUE           PROB
Moran's I (error)             -0.0238       1.270          0.2042
Lagrange Multiplier (lag)         1         1.312          0.2520
Robust LM (lag)                   1         0.038          0.8459
Lagrange Multiplier (error)       1         2.459          0.1169
Robust LM (error)                 1         1.184          0.2764
Lagrange Multiplier (SARMA)       2         2.497          0.2870

============================== END OF REPORT ================================
```

Regression report output

It is noted that:

1. The significant values with p-value less than 0.05 include:
   - "AMERICAN_INDIAN_ALASKA_NATIVE"
   - "INDIVIDUAL_INCOME_MEDIAN"
   - "WORKER_NOT_WFH"
   - "WORKER_VEHICLE_DROVE_ALONE"
   - "WORKER_VEHICLE_CARPOOL"
   - "WORKER_PUBLIC_TRANSPORTATION"
   - "POVERTY"
2. The model R-square is 0.7514, and its adjusted R-square is 0.6605

The fact that median income and poverty are associated with the number of bike stations seems to be interesting; this could be an indicator of more resources distributed to the rich.

Spatial Regression (Cont'd)

1. "Number of bike stations" versus "census features" (Cont'd)

(b) Regression using significant features identified:

```
# filter out significant coefficients (i.e. 8, 15, 16, 30, 31, 32, 33)
# get X and y
feature_cols = ["AMERICAN_INDIAN_ALASKA_NATIVE","INDIVIDUAL_INCOME_MEDIAN", "WORKER_NOT_WFH", 'WORKER_VEHICLE_DROVE_ALONE',
                'WORKER_VEHICLE_CARPOOL', 'WORKER_PUBLIC_TRANSPORTATION', 'POVERTY']
X = census_df[feature_cols].values

# standardization first
scaler = StandardScaler()
X = scaler.fit_transform(X)
y = census_df.bike_station_count.values

ols_spreg = spreg.OLS(y, X, w=W, spat_diag=True, moran = True)
print(ols_spreg.summary)
```

Python command

The spatial weights were calculated leveraging libpysal library with the below command:

```
# calculate weights
points = list(zip(census_df['INTPTLAT'], census_df['INTPTLON']))
W=libpysal.weights.DistanceBand(points,threshold=11.2,binary=False)
```

Spatial Regression (Cont'd)

1. "Number of bike stations" versus "census features" (Cont'd)

(b) Regression using significant features identified (Cont'd):

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  :     dep_var                Number of Observations:         128
Mean dependent var  :      0.7812                Number of Variables   :           8
S.D. dependent var  :      1.9398                Degrees of Freedom    :         120
R-squared           :      0.6012
Adjusted R-squared  :      0.5780
Sum squared residual:     190.553                F-statistic           :     25.8486
Sigma-square        :       1.588                Prob(F-statistic)     :   2.861e-21
S.E. of regression  :       1.260                Log likelihood        :    -207.090
Sigma-square ML     :       1.489                Akaike info criterion :     430.179
S.E of regression ML:      1.2201                Schwarz criterion     :     452.996

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT       0.7812500       0.1113812       7.0141994       0.0000000
               var_1      -0.2729384       0.1154025      -2.3650995       0.0196290
               var_2       0.2850052       0.1361439       2.0934120       0.0384183
               var_3       4.6204681       0.3745235      12.3369240       0.0000000
               var_4      -2.8472468       0.3110419      -9.1538999       0.0000000
               var_5      -0.4639421       0.1589958      -2.9179522       0.0042077
               var_6      -0.9081053       0.1912089      -4.7492833       0.0000057
               var_7      -0.8794293       0.3308050      -2.6584525       0.0089206
------------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER           8.148

TEST ON NORMALITY OF ERRORS
TEST                             DF        VALUE           PROB
Jarque-Bera                       2       265.890          0.0000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                             DF        VALUE           PROB
Breusch-Pagan test                7       193.137          0.0000
Koenker-Bassett test              7        46.407          0.0000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                           MI/DF       VALUE           PROB
Moran's I (error)             -0.0034        0.723          0.4698
Lagrange Multiplier (lag)         1          5.203          0.0225
Robust LM (lag)                   1          6.532          0.0106
Lagrange Multiplier (error)       1          0.049          0.8255
Robust LM (error)                 1          1.377          0.2405
Lagrange Multiplier (SARMA)       2          6.581          0.0372

================================ END OF REPORT =====================================
```

Regression report output

It is noted that:

1. All significant have a p-value less than 0.05:
   - "AMERICAN_INDIAN_ALASKA_NATIVE"
   - "INDIVIDUAL_INCOME_MEDIAN"
   - "WORKER_NOT_WFH"
   - "WORKER_VEHICLE_DROVE_ALONE"
   - "WORKER_VEHICLE_CARPOOL"
   - "WORKER_PUBLIC_TRANSPORTATION"
   - "POVERTY"
2. The model R-square is 0.6012, and its adjusted R-square is 0.5780

The result reinforced our previous finding: bike station counts may be associated with income level

Spatial Regression

2. “Average activity” versus “number of bike stations & census features”

(a) Regression using the full set of features:

```
# full features first
# get X and y (for full set)
feature_cols = ['POPULATION_DENSITY', "AGE_UNDER_17", "AGE_18_TO_64", "AGE_OVER_65", "MALE", 'WHITE',
                "BLACK_AFRICAN_AMERICAN", "AMERICAN_INDIAN_ALASKA_NATIVE", "ASIAN", "HAWAIIAN_PACIFIC",
                "HISPANIC_LATINO", "LANGUAGE_OTHER_THAN_ENGLISH", 'MEDIAN_HOUSE_VALUE',
                'COLLEGE_DEGREE_OR_ABOVE', "INDIVIDUAL_INCOME_MEDIAN", "WORKER_NOT_WFH",
                "WORKER_DEPART_0000_0559", "WORKER_DEPART_0600_0729", "WORKER_DEPART_0730_0859", "WORKER_DEPART_0900_2359",
                "TRAVEL_TIME_TO_WORK_LESS_THAN_10_MIN", 'TRAVEL_TIME_TO_WORK_10_TO_14_MIN',
                'TRAVEL_TIME_TO_WORK_15_TO_19_MIN', 'TRAVEL_TIME_TO_WORK_20_TO_24_MIN',
                'TRAVEL_TIME_TO_WORK_25_TO_29_MIN', 'TRAVEL_TIME_TO_WORK_30_TO_34_MIN',
                'TRAVEL_TIME_TO_WORK_35_TO_44_MIN', 'TRAVEL_TIME_TO_WORK_45_TO_59_MIN',
                'TRAVEL_TIME_TO_WORK_MORE_THAN_60_MIN', 'WORKER_VEHICLE_DROVE_ALONE',
                'WORKER_VEHICLE_CARPOOL', 'WORKER_PUBLIC_TRANSPORTATION', 'POVERTY', 'DISABILITY', 'bike_station_count']
X = census_df[feature_cols].values

# standardization first
scaler = StandardScaler()
X = scaler.fit_transform(X)
avg_count = np.array(census_df['total_activity_count']/census_df['bike_station_count'])
avg_count = np.nan_to_num(avg_count, nan=0)
y = avg_count

ols_spreg = spreg.OLS(y, X, w=W, spat_diag=True, moran = True)
print(ols_spreg.summary)
```

Python command

The spatial weights were calculated leveraging libpysal library with the below command:

```
# calculate weights
points = list(zip(census_df['INTPTLAT'], census_df['INTPTLON']))
W=libpysal.weights.DistanceBand(points,threshold=11.2,binary=False)
```

Spatial Regression (Cont'd)

2. "Average activity" versus "number of bike stations & census features" (Cont'd)

(a) Regression using the full set of features (Cont'd):

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  :     dep_var                Number of Observations:     128
Mean dependent var  :    607.0060                Number of Variables   :      36
S.D. dependent var  :   1036.6933                Degrees of Freedom    :      92
R-squared           :      0.5012
Adjusted R-squared  :      0.3114
Sum squared residual:68086792.025                F-statistic           :    2.6408
Sigma-square        :  740073.826                Prob(F-statistic)     : 0.0001141
S.E. of regression  :     860.275                Log likelihood        : -1025.417
Sigma-square ML     :  531928.063                Akaike info criterion :  2122.834
S.E of regression ML:    729.3340                Schwarz criterion     :  2225.507
```

Variable	Coefficient	Std.Error	t-Statistic	Probability
CONSTANT	607.0060016	76.0383243	7.9828956	0.0000000
var_1	122.1682042	122.2201254	0.9995752	0.3201380
var_2	-449.0132582	327.0392352	-1.3729645	0.1731007
var_3	-577.0488856	743.7778195	-0.7758350	0.4398363
var_4	-411.9655874	342.9725735	-1.2011619	0.2327717
var_5	30.3005510	150.8201316	0.2009052	0.8412162
var_6	633.1532337	978.0313738	0.6473752	0.5190013
var_7	377.1609049	855.1653527	0.4410386	0.6602198
var_8	59.9753681	93.2118543	0.6434307	0.5215453
var_9	217.2920796	266.0239056	0.8168141	0.4161448
var_10	-18.4884863	90.8830615	-0.2034316	0.8392470
var_11	-41.5969199	116.1168267	-0.3582334	0.7209894
var_12	-34.7404485	208.8880607	-0.1663113	0.8682769
var_13	87.2100541	153.4148093	0.5684592	0.5711082
var_14	-4.5709275	248.2355632	-0.0184137	0.9853487
var_15	-145.8651694	195.8064289	-0.7449458	0.4582033
var_16	-555.7375132	620.5115903	-0.8956118	0.3727975
var_17	2641.9129966	4864.2747928	0.5431258	0.5883567
var_18	4635.9436025	9136.4151668	0.5074139	0.6130787
var_19	5781.7521943	10859.5068420	0.5324139	0.5957229
var_20	5577.1233618	11154.1533239	0.5000042	0.6182660
var_21	-2715.5256201	5684.3800257	-0.4777171	0.6339846
var_22	-2609.7096376	5296.4271663	-0.4927302	0.6233772
var_23	-3133.2102367	6164.3889259	-0.5082759	0.6124765
var_24	-3188.6486680	6399.2467034	-0.4982850	0.6194723
var_25	-1676.0938518	3171.2139663	-0.5285338	0.5984016
var_26	-3060.9321950	6123.7905634	-0.4998427	0.6183792
var_27	-1392.6070048	2983.3635666	-0.4667909	0.6417530
var_28	-1278.6303427	2473.7113771	-0.5168874	0.6064751
var_29	-2085.5050739	3890.1338187	-0.5361011	0.5931825
var_30	-107.7168610	381.2662760	-0.2825240	0.7781764
var_31	-49.4605995	130.5409812	-0.3788894	0.7056430
var_32	175.8928662	188.6280965	0.9324850	0.3535267
var_33	400.5929242	346.4912296	1.1561416	0.2506167
var_34	-9.2028428	101.4968474	-0.0906712	0.9279510
var_35	349.9528519	152.4986211	2.2947935	0.0240191

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER 986.015

TEST ON NORMALITY OF ERRORS

TEST	DF	VALUE	PROB
Jarque-Bera	2	171.539	0.0000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS

TEST	DF	VALUE	PROB
Breusch-Pagan test	35	131.936	0.0000
Koenker-Bassett test	35	38.615	0.3095

DIAGNOSTICS FOR SPATIAL DEPENDENCE

TEST	MI/DF	VALUE	PROB
Moran's I (error)	-0.0126	0.092	0.9267
Lagrange Multiplier (lag)	1	3.094	0.0786
Robust LM (lag)	1	9.460	0.0021
Lagrange Multiplier (error)	1	0.692	0.4056
Robust LM (error)	1	7.058	0.0079
Lagrange Multiplier (SARMA)	2	10.152	0.0062

================================ END OF REPORT ====================================

Regression report output

It is noted that:

1. The number of bike stations is dominating the model, and no other features are considered significant
2. The R-square is 0.5012, and the adjusted R-square is 0.3114.

This model is unsatisfying as we can see that it is dominated by the number of bike stations and does not have explanatory power from other features

Spatial Regression (Cont'd)

2. "Average activity" versus "number of bike stations & census features" (Cont'd)

(b) Regression after removing the dominant feature (bike station count):

```
# remove bike station counts
# get X and y
feature_cols = ['POPULATION_DENSITY', "AGE_UNDER_17", "AGE_18_TO_64", "AGE_OVER_65", "MALE", 'WHITE',
                "BLACK_AFRICAN_AMERICAN", "AMERICAN_INDIAN_ALASKA_NATIVE", "ASIAN", "HAWAIIAN_PACIFIC",
                "HISPANIC_LATINO", "LANGUAGE_OTHER_THAN_ENGLISH", 'MEDIAN_HOUSE_VALUE',
                'COLLEGE_DEGREE_OR_ABOVE', "INDIVIDUAL_INCOME_MEDIAN", "WORKER_NOT_WFH",
                "WORKER_DEPART_0000_0559", "WORKER_DEPART_0600_0729", "WORKER_DEPART_0730_0859", "WORKER_DEPART_0900_2359",
                "TRAVEL_TIME_TO_WORK_LESS_THAN_10_MIN", 'TRAVEL_TIME_TO_WORK_10_TO_14_MIN',
                'TRAVEL_TIME_TO_WORK_15_TO_19_MIN', 'TRAVEL_TIME_TO_WORK_20_TO_24_MIN',
                'TRAVEL_TIME_TO_WORK_25_TO_29_MIN', 'TRAVEL_TIME_TO_WORK_30_TO_34_MIN',
                'TRAVEL_TIME_TO_WORK_35_TO_44_MIN', 'TRAVEL_TIME_TO_WORK_45_TO_59_MIN',
                'TRAVEL_TIME_TO_WORK_MORE_THAN_60_MIN', 'WORKER_VEHICLE_DROVE_ALONE',
                'WORKER_VEHICLE_CARPOOL', 'WORKER_PUBLIC_TRANSPORTATION', 'POVERTY', 'DISABILITY']
X = census_df[feature_cols].values

# standardization first
scaler = StandardScaler()
X = scaler.fit_transform(X)
avg_count = np.array(census_df['total_activity_count']/census_df['bike_station_count'])
avg_count = np.nan_to_num(avg_count, nan=0)
y = avg_count

ols_spreg = spreg.OLS(y, X, w=W, spat_diag=True, moran = True)
print(ols_spreg.summary)
```

Python command

The spatial weights were calculated leveraging libpysal library with the below command:

```
# calculate weights
points = list(zip(census_df['INTPTLAT'], census_df['INTPTLON']))
W=libpysal.weights.DistanceBand(points,threshold=11.2,binary=False)
```

Spatial Regression (Cont'd)

2. "Average activity" versus "number of bike stations & census features" (Cont'd)

(b) Regression after removing the dominant feature (bike station count) (Cont'd):

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  :     dep_var          Number of Observations:     128
Mean dependent var  :    607.0060          Number of Variables   :      35
S.D. dependent var  :   1036.6933          Degrees of Freedom    :      93
R-squared           :      0.4726
Adjusted R-squared  :      0.2798
Sum squared residual:71984078.099          F-statistic           :      2.4512
Sigma-square        :  774022.345          Prob(F-statistic)     :   0.0003678
S.E. of regression  :     879.785          Log likelihood        :   -1028.979
Sigma-square ML     :  562375.610          Akaike info criterion :    2127.959
S.E of regression ML:    749.9171          Schwarz criterion     :    2227.780

------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------
            CONSTANT     607.0060016      77.7627775       7.8058683       0.0000000
               var_1     102.3000891     124.6779171       0.8205149       0.4140208
               var_2    -520.0306791     332.9552563      -1.5618636       0.1217167
               var_3    -667.3549065     759.5803523      -0.8785837       0.3818914
               var_4    -481.9204170     349.3625385      -1.3794279       0.1710709
               var_5      -3.1605995     153.5180002      -0.0205878       0.9836186
               var_6     790.9464053     997.7368446       0.7927405       0.4299467
               var_7     615.0348087     868.1107614       0.7084750       0.4804224
               var_8       6.0431981      92.2460029       0.0655118       0.9479072
               var_9     284.7999843     270.3884220       1.0532995       0.2949327
              var_10     -13.2275577      92.9145990      -0.1423625       0.8871016
              var_11     -39.4747976     118.7464437      -0.3324293       0.7403132
              var_12     -33.7235051     213.6248975      -0.1578632       0.8749069
              var_13      82.8976329     156.8822925       0.5284066       0.5984759
              var_14    -111.0995217     249.3866747      -0.4454910       0.6570003
              var_15      47.8082431     180.6912265       0.2645853       0.7919144
              var_16     331.7896880     496.2402986       0.6686069       0.5054025
              var_17    1814.3020526    4960.8987979       0.3657204       0.7154042
              var_18    3070.7804387    9317.5456855       0.3295697       0.7424663
              var_19    3822.1989731   11071.3993831       0.3452318       0.7306997
              var_20    3594.8864896   11372.8584588       0.3160935       0.7526399
              var_21   -1637.4878318    5793.4084571      -0.2826467       0.7780758
              var_22   -1662.8220274    5400.0809275      -0.3079254       0.7588277
              var_23   -2097.1607134    6287.2587206      -0.3335572       0.7394646
              var_24   -2087.7859724    6525.9585750      -0.3199202       0.7497464
              var_25   -1132.9098135    3234.0862620      -0.3503029       0.7269035
              var_26   -1994.5419281    6244.6127800      -0.3194020       0.7501380
              var_27    -879.2689428    3042.4336472      -0.2890018       0.7732232
              var_28    -860.7678921    2522.9487887      -0.3411753       0.7337412
              var_29   -1425.1785316    3967.4585548      -0.3592170       0.7202471
              var_30    -624.7242797     314.5560340      -1.9860509       0.0499719
              var_31    -148.7146006     125.9607612      -1.1806423       0.2407557
              var_32      45.6133433     183.9614666       0.2479505       0.8047193
              var_33     237.2821902     346.7945766       0.6842154       0.4955406
              var_34       1.9135037     103.6803796       0.0184558       0.9853148
------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER         982.124

TEST ON NORMALITY OF ERRORS
TEST                             DF        VALUE           PROB
Jarque-Bera                       2        138.116         0.0000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                             DF        VALUE           PROB
Breusch-Pagan test               34        112.518         0.0000
Koenker-Bassett test             34         36.062         0.3723

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                           MI/DF       VALUE           PROB
Moran's I (error)             -0.0103      0.394           0.6936
Lagrange Multiplier (lag)         1        2.616           0.1058
Robust LM (lag)                   1        7.768           0.0053
Lagrange Multiplier (error)       1        0.459           0.4979
Robust LM (error)                 1        5.611           0.0178
Lagrange Multiplier (SARMA)       2        8.227           0.0163

================================ END OF REPORT ====================================
```

Regression report output

It is noted that:

1. No features can be considered as significant
2. The R-square is 0.4726, and the adjusted R-square is 0.2798

This model is unsatisfying as we could not find any potential insights

Elastic Net

1. "Number of bike stations" versus "census features"

```
from scipy.stats import t
# full features first
# get X and y (for full set)
feature_cols = ['POPULATION_DENSITY', "AGE_UNDER_17", "AGE_18_TO_64", "AGE_OVER_65", "MALE", 'WHITE',
                "BLACK_AFRICAN_AMERICAN", "AMERICAN_INDIAN_ALASKA_NATIVE", "ASIAN", "HAWAIIAN_PACIFIC",
                "HISPANIC_LATINO", "LANGUAGE_OTHER_THAN_ENGLISH", 'MEDIAN_HOUSE_VALUE',
                'COLLEGE_DEGREE_OR_ABOVE', "INDIVIDUAL_INCOME_MEDIAN", "WORKER_NOT_WFH",
                "WORKER_DEPART_0000_0559", "WORKER_DEPART_0600_0729", "WORKER_DEPART_0730_0859", "WORKER_DEPART_0900_2359",
                "TRAVEL_TIME_TO_WORK_LESS_THAN_10_MIN", 'TRAVEL_TIME_TO_WORK_10_TO_14_MIN',
                'TRAVEL_TIME_TO_WORK_15_TO_19_MIN', 'TRAVEL_TIME_TO_WORK_20_TO_24_MIN',
                'TRAVEL_TIME_TO_WORK_25_TO_29_MIN', 'TRAVEL_TIME_TO_WORK_30_TO_34_MIN',
                'TRAVEL_TIME_TO_WORK_35_TO_44_MIN', 'TRAVEL_TIME_TO_WORK_45_TO_59_MIN',
                'TRAVEL_TIME_TO_WORK_MORE_THAN_60_MIN', 'WORKER_VEHICLE_DROVE_ALONE',
                'WORKER_VEHICLE_CARPOOL', 'WORKER_PUBLIC_TRANSPORTATION', 'POVERTY', 'DISABILITY']
X = census_df[feature_cols].values

# standardization first
scaler = StandardScaler()
X = scaler.fit_transform(X)
y = census_df.bike_station_count.values

# create elastic net model
enet_model = ElasticNet(alpha=0.1, l1_ratio=0.4, fit_intercept=False)
enet_results = enet_model.fit(X, y)
```

```
# calculate R-squared
r_squared = enet_model.score(X, y)

# calculate p-values
n = len(y)
p = X.shape[1]
rss = np.sum((y - enet_model.predict(X))**2)
se = np.sqrt(rss / (n - p) / np.diag(np.dot(X.T, X)))

# Calculate t-values and p-values
t_values = enet_model.coef_ / se
p_values = 2 * t.cdf(-np.abs(t_values), n - p)

# display results
print("Elastic Net Results:")
print("R-squared: {:.3f}".format(r_squared))
print('{:<40} {:10} {:10} {:10}'.format('feature', 'coefficient','t-value','p-value'))
for feature, coef, t, p in zip(feature_cols, enet_results.coef_, t_values, p_values):
    print('{:<40} {:10} {:10} {:10}'.format(feature, np.round(coef, 6), np.round(t, 4), np.round(p, 4)))
```

Python command

Elastic Net (Cont'd)

1. "Number of bike stations" versus "census features" (Cont'd)

Elastic Net Results:
R-squared: 0.418

feature	coefficient	t-value	p-value
POPULATION_DENSITY	-0.060146	-0.3956	0.6933
AGE_UNDER_17	-0.451344	-2.9684	0.0038
AGE_18_TO_64	0.086754	0.5706	0.5697
AGE_OVER_65	-0.316266	-2.08	0.0402
MALE	-0.100338	-0.6599	0.5109
WHITE	-0.095409	-0.6275	0.5319
BLACK_AFRICAN_AMERICAN	0.144822	0.9525	0.3433
AMERICAN_INDIAN_ALASKA_NATIVE	-0.167312	-1.1004	0.274
ASIAN	0.060513	0.398	0.6915
HAWAIIAN_PACIFIC	-0.013112	-0.0862	0.9315
HISPANIC_LATINO	-0.0	-0.0	1.0
LANGUAGE_OTHER_THAN_ENGLISH	-0.0	-0.0	1.0
MEDIAN_HOUSE_VALUE	0.088407	0.5814	0.5623
COLLEGE_DEGREE_OR_ABOVE	-0.0	-0.0	1.0
INDIVIDUAL_INCOME_MEDIAN	0.460763	3.0303	0.0032
WORKER_NOT_WFH	1.016425	6.6847	0.0
WORKER_DEPART_0000_0559	0.087847	0.5777	0.5648
WORKER_DEPART_0600_0729	0.011752	0.0773	0.9386
WORKER_DEPART_0730_0859	-0.187438	-1.2327	0.2208
WORKER_DEPART_0900_2359	-0.103635	-0.6816	0.4972
TRAVEL_TIME_TO_WORK_LESS_THAN_10_MIN	0.76367	5.0224	0.0
TRAVEL_TIME_TO_WORK_10_TO_14_MIN	0.21085	1.3867	0.1688
TRAVEL_TIME_TO_WORK_15_TO_19_MIN	-0.185272	-1.2185	0.2261
TRAVEL_TIME_TO_WORK_20_TO_24_MIN	-0.034039	-0.2239	0.8233
TRAVEL_TIME_TO_WORK_25_TO_29_MIN	-0.027557	-0.1812	0.8566
TRAVEL_TIME_TO_WORK_30_TO_34_MIN	-0.0	-0.0	1.0
TRAVEL_TIME_TO_WORK_35_TO_44_MIN	-0.048671	-0.3201	0.7496
TRAVEL_TIME_TO_WORK_45_TO_59_MIN	-0.031371	-0.2063	0.837
TRAVEL_TIME_TO_WORK_MORE_THAN_60_MIN	-0.0	-0.0	1.0
WORKER_VEHICLE_DROVE_ALONE	-0.627428	-4.1264	0.0001
WORKER_VEHICLE_CARPOOL	-0.185499	-1.22	0.2255
WORKER_PUBLIC_TRANSPORTATION	0.0	0.0	1.0
POVERTY	-0.0	-0.0	1.0
DISABILITY	0.0	0.0	1.0

It is noted that:

1. The significant values with p-value less than 0.05 include
   - "AGE_UNDER_17"
   - "AGE_OVER_65"
   - "INDIVIDUAL_INCOME_MEDIAN"
   - "WORKER_NOT_WFH"
   - "TRAVEL_TIME_TO_WORK_LESS_THAN_10_MIN"
   - "WORKER_VEHICLE_DROVE_ALONE"

2. The R-square is 0.418

The fact that median income is associated with the number of bike stations seems to be interesting (similar like the result we got from spatial regression); this could be an indicator of more resources distributed to the rich.

Elastic Net (Cont'd)

2. "Average activity" versus "number of bike stations & census features"

```python
from scipy.stats import t
# full features first
# get X and y (for full set)
feature_cols = ['POPULATION_DENSITY', "AGE_UNDER_17", "AGE_18_TO_64", "AGE_OVER_65", "MALE", 'WHITE',
                "BLACK_AFRICAN_AMERICAN", "AMERICAN_INDIAN_ALASKA_NATIVE", "ASIAN", "HAWAIIAN_PACIFIC",
                "HISPANIC_LATINO", "LANGUAGE_OTHER_THAN_ENGLISH", 'MEDIAN_HOUSE_VALUE',
                'COLLEGE_DEGREE_OR_ABOVE', "INDIVIDUAL_INCOME_MEDIAN", "WORKER_NOT_WFH",
                "WORKER_DEPART_0000_0559", "WORKER_DEPART_0600_0729", "WORKER_DEPART_0730_0859", "WORKER_DEPART_0900_2359",
                "TRAVEL_TIME_TO_WORK_LESS_THAN_10_MIN", 'TRAVEL_TIME_TO_WORK_10_TO_14_MIN',
                'TRAVEL_TIME_TO_WORK_15_TO_19_MIN', 'TRAVEL_TIME_TO_WORK_20_TO_24_MIN',
                'TRAVEL_TIME_TO_WORK_25_TO_29_MIN', 'TRAVEL_TIME_TO_WORK_30_TO_34_MIN',
                'TRAVEL_TIME_TO_WORK_35_TO_44_MIN', 'TRAVEL_TIME_TO_WORK_45_TO_59_MIN',
                'TRAVEL_TIME_TO_WORK_MORE_THAN_60_MIN', 'WORKER_VEHICLE_DROVE_ALONE',
                'WORKER_VEHICLE_CARPOOL', 'WORKER_PUBLIC_TRANSPORTATION', 'POVERTY', 'DISABILITY', 'bike_station_count']
X = census_df[feature_cols].values

# standardization first
scaler = StandardScaler()
X = scaler.fit_transform(X)
avg_count = np.array(census_df['total_activity_count']/census_df['bike_station_count'])
avg_count = np.nan_to_num(avg_count, nan=0)
y = avg_count

# create elastic net model
enet_model = ElasticNet(alpha=0.01, l1_ratio=0.5, fit_intercept=False)
enet_results = enet_model.fit(X, y)
```

```python
# calculate R-squared
r_squared = enet_model.score(X, y)

# calculate p-values
n = len(y)
p = X.shape[1]
rss = np.sum((y - enet_model.predict(X))**2)
se = np.sqrt(rss / (n - p) / np.diag(np.dot(X.T, X)))

# Calculate t-values and p-values
t_values = enet_model.coef_ / se
p_values = 2 * t.cdf(-np.abs(t_values), n - p)

# display results
print("Elastic Net Results:")
print("R-squared: {:.3f}".format(r_squared))
print('{:<40} {:10} {:10} {:10}'.format('feature', 'coefficient','t-value','p-value'))
for feature, coef, t, p in zip(feature_cols, enet_results.coef_, t_values, p_values):
    print('{:<40} {:10} {:10} {:10}'.format(feature, np.round(coef, 6), np.round(t, 4), np.round(p, 4)))
```

Python command

Elastic Net (Cont'd)

2. "Average activity" versus "number of bike stations & census features" (Cont'd)

Elastic Net Results:
R-squared: 0.153

feature	coefficient	t-value	p-value
POPULATION_DENSITY	115.78167	1.1747	0.2431
AGE_UNDER_17	-299.768584	-3.0413	0.0031
AGE_18_TO_64	-215.43079	-2.1856	0.0314
AGE_OVER_65	-236.847874	-2.4029	0.0182
MALE	2.219528	0.0225	0.9821
WHITE	164.121597	1.6651	0.0993
BLACK_AFRICAN_AMERICAN	-6.743297	-0.0684	0.9456
AMERICAN_INDIAN_ALASKA_NATIVE	50.776635	0.5152	0.6077
ASIAN	135.577029	1.3755	0.1723
HAWAIIAN_PACIFIC	-25.100209	-0.2547	0.7996
HISPANIC_LATINO	-41.454163	-0.4206	0.675
LANGUAGE_OTHER_THAN_ENGLISH	-31.761178	-0.3222	0.748
MEDIAN_HOUSE_VALUE	84.11886	0.8534	0.3956
COLLEGE_DEGREE_OR_ABOVE	-53.666242	-0.5445	0.5874
INDIVIDUAL_INCOME_MEDIAN	-122.545193	-1.2433	0.2169
WORKER_NOT_WFH	-368.261003	-3.7362	0.0003
WORKER_DEPART_0000_0559	137.880158	1.3989	0.1652
WORKER_DEPART_0600_0729	-1.187196	-0.012	0.9904
WORKER_DEPART_0730_0859	254.330274	2.5803	0.0114
WORKER_DEPART_0900_2359	-94.897069	-0.9628	0.3382
TRAVEL_TIME_TO_WORK_LESS_THAN_10_MIN	156.509251	1.5879	0.1157
TRAVEL_TIME_TO_WORK_10_TO_14_MIN	84.016051	0.8524	0.3962
TRAVEL_TIME_TO_WORK_15_TO_19_MIN	4.830301	0.049	0.961
TRAVEL_TIME_TO_WORK_20_TO_24_MIN	65.000056	0.6595	0.5112
TRAVEL_TIME_TO_WORK_25_TO_29_MIN	-60.650369	-0.6153	0.5398
TRAVEL_TIME_TO_WORK_30_TO_34_MIN	52.584557	0.5335	0.595
TRAVEL_TIME_TO_WORK_35_TO_44_MIN	125.948688	1.2778	0.2045
TRAVEL_TIME_TO_WORK_45_TO_59_MIN	-17.200869	-0.1745	0.8618
TRAVEL_TIME_TO_WORK_MORE_THAN_60_MIN	-109.435743	-1.1103	0.2697
WORKER_VEHICLE_DROVE_ALONE	-148.339195	-1.505	0.1357
WORKER_VEHICLE_CARPOOL	-70.122997	-0.7114	0.4786
WORKER_PUBLIC_TRANSPORTATION	146.631791	1.4876	0.1402
POVERTY	296.14533	3.0045	0.0034
DISABILITY	-6.864234	-0.0696	0.9446
bike_station_count	324.517424	3.2924	0.0014

It is noted that:

1. The significant values with p-value less than 0.05 include
   - "AGE_UNDER_17"
   - "AGE_18_TO_64
   - "AGE_OVER_65"
   - "WORKER_NOT_WFH"
   - "WORDER_DEPART_0730_0859"
   - "POVERTY"
   - "BIKE_STATION_COUNT"

2. The R-square is 0.418

The fact that poverty is associated with the average activity seems to be interesting. It indicates that in less rich areas, the bikes are utilized more often. This could be because of the population there have higher demands for the bikes as bikes are more affordable for them.