Project

Sankalp Devasthali

Due Dec 16

1. Premise

The Pittsburgh region is served by a combined sewer, which collects, conveys, and partially treats both stormwater and wastewater in the same infrastructure system. During periods of heavy rainfall, stormwater can exceed the capacity of combined sewers, which causes overflow into nearby rivers and streams. While these combined sewer overflows or CSO's mitigate upstream flooding, they release untreated wastewater into receiving water bodies. CSO's are regulated by state and federal authorities, thus cost effective strategies to manage CSO's are important for local municipalities.

Improvements to a green infrastructure - pipes, pumps, storage, and treatment facilities - can increase the capacity of the collection system to accommodate more severe wet weather events. Conversely, a green infrastructure includes features that reduce the stormwater entering the collection system by temporarily retaining or diverting stormwater. Types of green infrastructure vary from completely natural systems, such as converting a parking lot to a park, to single purpose engineered systems, such as pervious paving.

While gray infrastructure strategies involve modifying public property, many green infrastructure strategies involve modifying private property. In particular, both rain gardens and trees involve retrofitting exterior space on private property. There are advantages and disadvantages to these three uses of outdoor space: open space ("grassy yard"), trees, and raingardens. For each of these uses of outdoor spaces, property owners will value differently effects on aesthetics, environmental impacts, maintenance, and outdoor uses (e.g., recreation is feasible on open space but not on a rain garden). Thus, it is important to understand how property owners value these amenities before considering rain gardens and trees as a stormwater management strategy.

2. Importing the data

I am getting my data from the amazon server provided to us https://s3.amazonaws.com/aws-website-programminginrforanalytics-tbal0/data/sales_county.csv (https://s3.amazonaws.com/aws-website-programminginrforanalytics-tbal0/data/sales_county.csv)
The files I am using are-:
1. sales_county.csv
2. assesment_city.csv
3. land_use_city.csv
4. sewersheds.csv

3. Description of information provided

A breif look at the dataset that I will be using to predict the rebate value and the effects of our strategy on the CSO values-:
(1) Sales prices for residential properties in Allegheny County

(2) Descriptions of residential properties in the City of Pittsburgh

(3) Land use data for parcels in the City

(4) Sewershed locations for parcels in the City

A quick explanation of the various columns in the given datas-

Assesment_City

1. PARID - Parcel identifier, ie, a plot of land uniquely identified by a code.
2. PROPERTYCITY - The city in which the property exists
3. MUNIDESC - Property municipality
4. USEDESC - Approved property use
5. LOTAREA - Lot or parcel area in Square Feet
6. HOMESTEADFLAG - Indicator for whether the property owner occupies the property.The assumption here is that a property can be occupied by owner (HOM) or renter only.
7. CONDITIONDESC - Property condition as assessed by Allegheny County
8. BEDROOMS - Number of bedrooms
9. FULLBATHS - Number of full bathrooms
10. HALFBATHS - Number of half bathrooms
11. FINISHEDLIVINGAREA - Square feet of living area
12. ZIP_CODE - Property zip code
13. census.block.group - Property Census block group
14. neighborhood - Property neighborhood

Sales_County

1. PARID - Parcel identifier, ie, a plot of land uniquely identified by a code.
2. SALEDATE - Date of sale of given property
3. SALEDESC - Type of sale of given property
4. PRICE - Price paid

Land_Use_City

1. PARID - Parcel identifier, ie, a plot of land uniquely identified by a code.
2. LUnew - Type of land use. “bldgs” indicates land occupied by buildings. “impervious” means impervious cover where water does not penetrate into the land. “trees” and “open.space” are explain the presence of trees on the property
3. sqft - Square feet of indicated land use

Sewersheds

1. PARID - Parcel identifier, ie, a plot of land uniquely identified by a code.
2. sewershed - Sewershed in which property is located
3. CSOperInfl - Ratio of combined sewer overflow reduced for every unit of runoff reduced at the surface

4. Tidying the data

For Assesment_city

In the assesment_city data, it can be seen that columns “Hood” and “neighborhood” are copies. Similarly for GEOID10

and census.block.group, we can see that they are identical.
So we can remove either of the pairs. I am removing “Hood” and “GEOID10” from my table.

The HOMESTEADFLAG stands for whether a home is occupied by it’s owner or if it has been rented out, thus, we create a new column “owner.rental” with the respective values.
Now, having two separate columns for Bathrooms doesn’t make enough sense. Let us combine the two and call it *bathrooms*. A higher weightage is give to the FULLBATHS value as compared to the HALFBATHS value.
Also, we will now filter data to only take those values which belong to “residential” properties, ie, remove values related to corporation as given in the *OWNERDESC*.
Finally, we will remove values which don’t belong to a household as given in the *USERDESC* column.

It doesn’t make sense to have fields where the values of FINISHEDLIVINGAREA and LOTAREA are not recorded or nill, hence we filter those values in our assesment_city data.

Sales_County

For the sales_county data, we only want valid sales where the price paid was significant and hence we filter on the basis of *SALESDESC* to only take values which pertain to a valid sale.

Finally, we drop all the irrelevant columns as they don’t have any effect on the price of a plot. Columns removed are - HOMESTEADFLAG, FULLBATHS, HALFBATHS, HOOD, OWNERDESC,SCHOOLDESC, GEOID10, census.block.group.

Merging Data

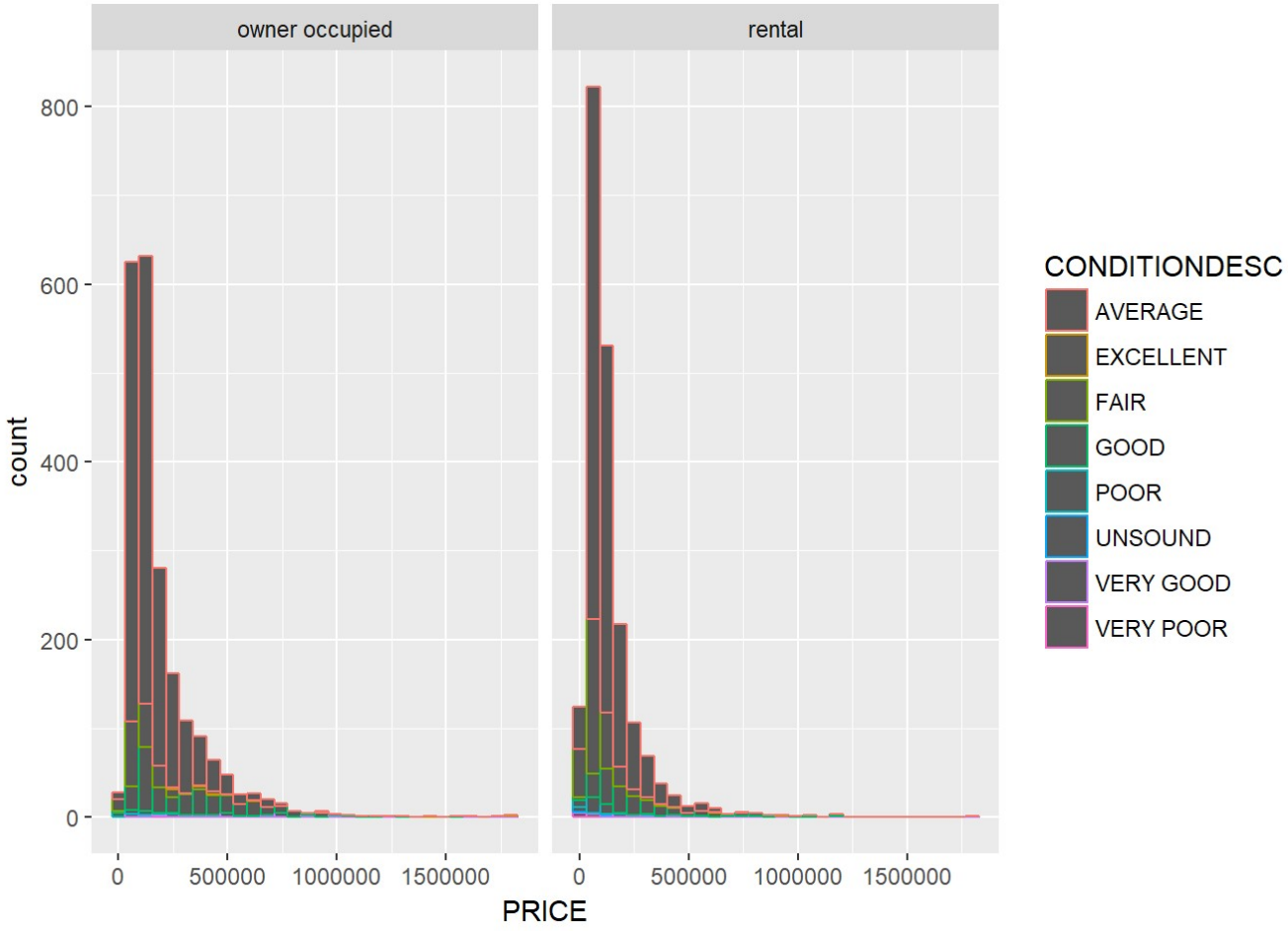
Now we merge data from the assesment_city dataset and sales_county dataset with land_use_city dataset. We can see that the common ID is the PARID.

Further Tidying

Some of the Properties have multiple *SALEDATEs*. We are only concerned about the latest one. Thus, we filter the data and keep only the latest saledates.

Finally, let us check whether there is a significant difference in the prices of houses for rental and for houses which the owners are the occupants

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



It can be safely said that the spread of houses is similar for either type - renter or owner occupied. There is similarity in the prices and the conditions and the skew is on the left side.

The questions we want to answer are whether the residents of Pittsburgh favour open spaces or trees. Since we have data about the usage of trees and open spaces (in square feet) for various properties, let us spread the data in order to regress.

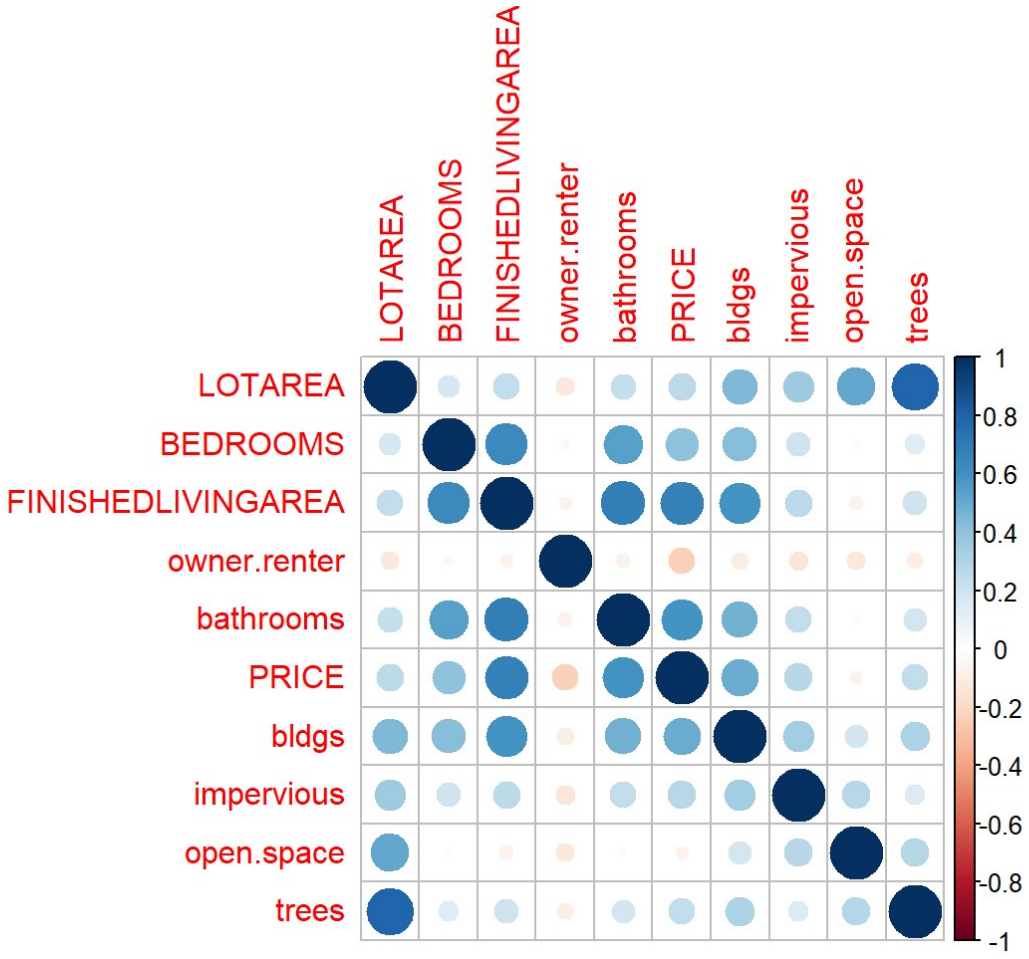
The final dataset that we use to answer our questions is -

```
##               PARID         SALEDATE              PROPERTYCITY
##  0001J00029000000:   1   Min.   :2012-01-03   CARNEGIE  :   6
##  0001N00053000000:   1   1st Qu.:2013-09-26   HOMESTEAD :  26
##  0001N00066000000:   1   Median :2015-04-27   PITTSBURGH:4124
##  0001N00165000000:   1   Mean   :2015-03-16
##  0001N00176000000:   1   3rd Qu.:2016-07-28
##  0001N00244000000:   1   Max.   :2017-10-05
##  (Other)         :4150
##                    MUNIDESC             USEDESC        LOTAREA
##  14th Ward - PITTSBURGH: 655   SINGLE FAMILY:3510   Min.   :  354
##  19th Ward - PITTSBURGH: 630   ROWHOUSE     : 331   1st Qu.: 2640
##  10th Ward - PITTSBURGH: 237   TWO FAMILY   : 251   Median : 3848
##  15th Ward - PITTSBURGH: 229   THREE FAMILY :  43   Mean   : 4690
##  20th Ward - PITTSBURGH: 222   FOUR FAMILY  :  16   3rd Qu.: 5614
##  27th Ward - PITTSBURGH: 215   MOBILE HOME  :   4   Max.   :67779
##  (Other)               :1968   (Other)      :   1
##     YEARBLT       CONDITIONDESC     BEDROOMS      FINISHEDLIVINGAREA
##  Min.   :1844   AVERAGE  :3011   Min.   : 1.000   Min.   : 399
##  1st Qu.:1906   FAIR     : 506   1st Qu.: 2.000   1st Qu.:1210
##  Median :1925   GOOD     : 492   Median : 3.000   Median :1496
##  Mean   :1926   VERY GOOD:  60   Mean   : 3.049   Mean   :1716
##  3rd Qu.:1945   POOR     :  56   3rd Qu.: 3.000   3rd Qu.:2000
##  Max.   :2016   EXCELLENT:  15   Max.   :12.000   Max.   :7276
##                 (Other)  :  16
##     ZIP_CODE                neighborhood          owner.renter
##  15217  : 562   Brookline          : 396   owner occupied:2162
##  15226  : 383   Squirrel Hill South: 279   rental        :1994
##  15212  : 376   Greenfield         : 197
##  15206  : 316   Carrick            : 187
##  15210  : 278   Beechview          : 168
##  15201  : 251   Squirrel Hill North: 154
##  (Other):1990   (Other)            :2775
##    bathrooms                            SALEDESC        PRICE
##  Min.   :0.500   VALID SALE                :2920   Min.   :   2000
##  1st Qu.:1.000   OTHER VALID               :1221   1st Qu.:  75000
##  Median :1.500   CHANGED AFTER SALE        :  15   Median : 115000
##  Mean   :1.548   BANK/FINANCIAL INSTITUTION:   0   Mean   : 168635
##  3rd Qu.:2.000   BUILDING NOT YET ASSESSED :   0   3rd Qu.: 195000
##  Max.   :6.500   CITY TREASURER SALE       :   0   Max.   :1800000
##                  (Other)                   :   0
##      bldgs          impervious       open.space          trees
##  Min.   :   0.0   Min.   :   0.0   Min.   :    0.0   Min.   :    0.0
```

```
##  1st Qu.: 716.4   1st Qu.: 573.2   1st Qu.:  105.1   1st Qu.:  267.5
##  Median : 949.6   Median : 902.1   Median :  471.9   Median :  914.8
##  Mean   :1050.3   Mean   :1011.9   Mean   :  826.1   Mean   : 1558.2
##  3rd Qu.:1267.4   3rd Qu.:1312.5   3rd Qu.: 1144.5   3rd Qu.: 1979.9
##  Max.   :4839.5   Max.   :5759.3   Max.   :10424.9   Max.   :58691.1
##
```

5. Plots

To see how the final data is correlated to each other, we shall use a CORRPLOT.



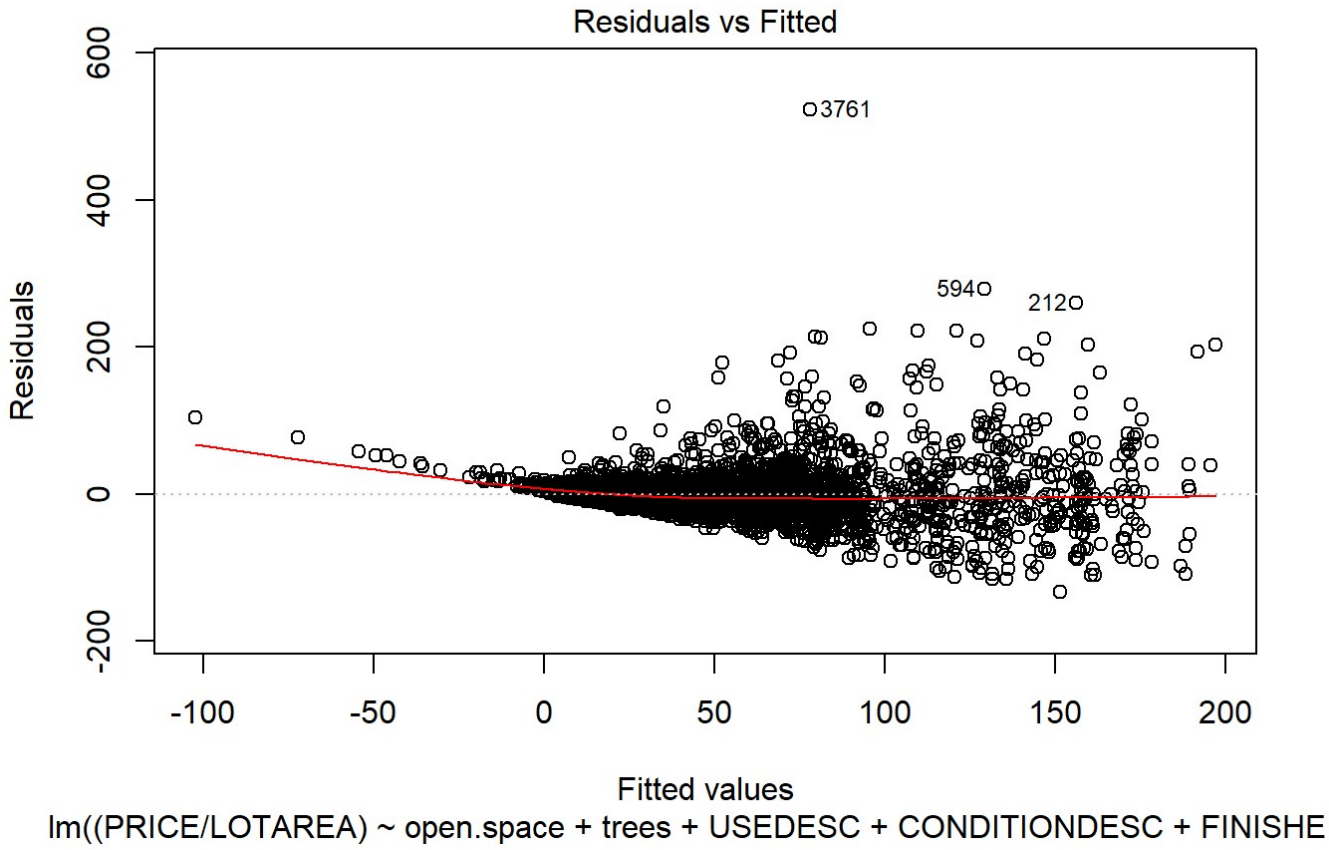
From the plot above, we can see that Price is correlated to lotarea, finishedlivingarea, bedrooms, bathrooms, bldgs and trees.

This may lead us to believe that our regression should include the aforementioned independent variables to get an estimate for the percieved value of trees and open space for residents of Pittsburgh. This is a wrong assumption.

BEDROOMS is correlated with FINISHEDLIVINGAREA and bldgs.

Thus, we can remove these estimators and still get a gainful regression with sufficient power to give us an answer to our questions. Removing unimportant variables and running the following regression, we get-:

```
regression <- lm(formula = (PRICE/LOTAREA) ~
                   open.space +
                   trees +
                   USEDESC +
                   CONDITIONDESC +
                   FINISHEDLIVINGAREA +
                   neighborhood
                   , data = merged.data)
plot(regression, which = 1)
```




```
##                 Estimate    Std. Error   t value     Pr(>|t|)
## (Intercept) 99.979180208 15.0987741446  6.621675 4.015639e-11
## open.space  -0.003662574  0.0005425711 -6.750404 1.683063e-11
## trees       -0.001734179  0.0002092199 -8.288787 1.537705e-16
```

6. Interpretation

We shall concern ourselves with the variables under consideration - trees and open.space.

trees - when controlling for the condition for the residential houses, having similar living areas and open spaces and

neighborhood, on an average, the value of a property per square feet decreases by $0.0017 for every increase in square feet of trees. This estimate is statistically significant.

open.space - when controlling for the condition for the residential houses, having similar living areas and tree count and neighborhood, on an average, the value of a property per square feet decreases by $0.0037 for every increase in square feet of open space. This estimate is statistically significant.

As can be seen, people don't value trees nor open space.

7. Estimating the rebate amount

The rebate will be calculated as compensation for the use of private property plus installation costs. As the perceived value of trees and open spaces is "negative" and "negative", respectively, we need to provide incentive to the household owners of Pittsburgh to plant trees in their yards. Following is a table which summarises the parameters to calculate rebate amount.

Amenity	Diameter	Installation cost
Trees	25 ft	$20 per tree
Rain gardens	NA	$6 per square foot

Rebate for trees

The coefficient for trees is -0.0017342. We multiply this with (pi x (diameter/2)^2) and add USD 20 to find the cost of putting a tree in a property. Thus, it is 20.8340625$

Rebate for open space

Since we don't have to incentivize converting open spaces to rainwater sheds, the rebate amount is simply 6$ per square foot.

8. Stormwater performance

The following are mean installation costs and stormwater performance data.

Amenity	Stormwater performance
Trees	1000 gallons of runoff reduced per tree
Rain gardens	3.5 gallons of runoff reduced per square foot

To find the most cost effective sewershed, we will merge sewersheds data with the current dataset on the common column "PARID". We then sum up the values of combined runoff reduced per square foot over all the sewerids.
The top 5 cost-effective sewerids are -

```
## # A tibble: 5 x 2
##          sewershed sum.per.cso
##              <fctr>       <dbl>
## 1              M-29   1049.2371
## 2 A-41/121H001-OF    661.7666
## 3              A-42    635.2303
## 4              A-22    513.7757
## 5              O-27    397.5002
```

9. Conclusion

We imported publicly available dataset for the city of Pittsburgh which included data about property sales, sewersheds, the usage of land, and plots in the city. We trimmed the data, stripping it of duplicate, redundant, unnecessary data. Thereafter, we answered 4 questions -
(1) How much do residential property owners value - if at all - exterior open space?
(2) How much do residential property owners value - if at all - trees?
(3) What is the mean rebate per square foot of rain garden and per tree the municipality should offer property owners to incentivize their installation?
(4) The five areas of the city (sewersheds) that are the most cost effective at reducing combined sewer overflows?

The people of Pittsburgh value trees and open spaces of their property negatively, hence we have to incentivise the inclusion of trees in their property. We found mean rebate values for planting trees and putting rainwater sheds. We also found the top 5 most effective sewersheds, which are-:
1. M-29
2. A-41/121H001-OF
3. A-42
4. A-22
5. O-27

REFERENCES

Blackhurst. MF. (2017). “Parcel Scale Green Infrastructure Siting and Cost Effectiveness Analysis.” http://sb.ucsur.pitt.edu/green-infrastructure/ (http://sb.ucsur.pitt.edu/green-infrastructure/)"

Allegheny County. Allegheny County Urban Tree Canopy. Division of Computer Services Geographic Information Systems Group, 2010. http://www.pasda.psu.edu/uci/MetadataDisplay.aspx?entry=PASDA&file=AlleghenyCountyUrbanTreeCanopy2010.xml&dataset=1203.com (http://www.pasda.psu.edu/uci/MetadataDisplay.aspx?entry=PASDA&file=AlleghenyCountyUrbanTreeCanopy2010.xml&dataset=1203.com)"

“Allegheny County Wooded Areas. Division of Computer Services Geographic Information Systems Group, 2011. http://www.pasda.psu.edu/uci/MetadataDisplay.aspx?entry=PASDA&file=AlleghenyCounty_WoodedAreas2011.xml&dataset=1228.com (http://www.pasda.psu.edu/uci/MetadataDisplay.aspx?entry=PASDA&file=AlleghenyCounty_WoodedAreas2011.xml&dataset=1228.com)”

“Allegheny County Property Assessments.” https://data.wprdc.org/dataset/property-assessments.com (https://data.wprdc.org/dataset/property-assessments.com)"

“Allegheny County Property Sale Transactions.” https://data.wprdc.org/dataset/real-estate-sales.com (https://data.wprdc.org/dataset/real-estate-sales.com)"

“City of Pittsburgh. Parcels. Geographic Data, 2015. http://pittsburghpa.gov/dcp/gis/gis-data-new.com (http://pittsburghpa.gov/dcp/gis/gis-data-new.com)”

“Street Curbs. Geographic Data, 2015. http://pittsburghpa.gov/dcp/gis/gis-data-new.com (http://pittsburghpa.gov/dcp/gis/gis-data-new.com)”

“PWSA (Pittsburgh Water and Sewer Authority). 2016. Sewershed Overview Map. http://www.arcgis.com/home/webmap/viewer.html?webmap=f96943c1e46e48dcad9abe5282bc58a8&extent=-80.2691,40.3363,-79.7621,40.5663 (http://www.arcgis.com/home/webmap/viewer.html?webmap=f96943c1e46e48dcad9abe5282bc58a8&extent=-80.2691,40.3363,-79.7621,40.5663)”

“tidyverse.org. http://tidyr.tidyverse.org/ (http://tidyr.tidyverse.org/)”
“rpubs.com. https://rpubs.com/bradleyboehmke/data_wrangling (https://rpubs.com/bradleyboehmke/data_wrangling)”