Allegheny County's relationship between economic performance and social equity, and the influence of walkability

Brent Ripperger

Carnegie Mellon University, Heinz College: 94-802 Geographic Information Systems

ABSTRACT:

This project explored the relationship between economic performance (EP), social equity (SE), and walkability (W). The purpose was to answer the question: does walkability explain why certain census tracts both have high EP and high SE while other tracts have high EP but low SE. The project analyzed census tracts inside Allegheny County using data from the US Census Bureau, Environmental Protection Agency (EPA), and Western Pennsylvania Regional Data Center (WPRDC). The hypothesis is that high walkability is associated with places of high EP and high SE, while low walkability is associated with places of high EP and low SE. However, the results show the opposite being true. This analysis found that SE is negatively correlated with EP $(pvalue =< 2e-16, Rsquared = 0.19)$. Similarly, walkability is negatively correlated with EP $(pvalue =< 2e-16, Rsquared = 0.25)$. Considering both SE and W were negatively correlated with EP, there becomes a need to understand the relationship between W and SE. This analysis found that walkability is positively correlated with SE $(pvalue = 0.000218, Rsquared = 0.034)$. While the R-squared value is small, the implications for this are interesting and surprising. The positive relationship between W and SE justifies the need for further analysis by policy makers. These findings could provide the rationale to improve the walkability of targeted census tracts, expecting to drive an increase in SE. While the project didn't prove that higher walkability explains why tracts have both high EP and high SE, the project was able to find a statistically significant, positive correlation between walkability and social equity.

INTRODUCTION:

During an analysis of walkable urban environments, Leinberger et al. found that "Walk Score by itself explains 67% of the increase in economic performance" (Leinberger, 2012). However, Leinberger et al. analysis was based solely on regionally significant, walkable urban environments. The boundaries of which weren't based on census tract data but specific criteria. Their results pose a new question, would the same correlation be found when analyzing census tract data? Will the walkability of census tracts be positively correlated with their EP? Additionally, Leinberger et al. analysis found that EP is negatively correlated with SE. Again, the question becomes, will the EP of census tracts be negatively correlated with their SE? In this project, the plan is to explore the relationship between EP, SE, and walkability by census tracts. Specifically, to answer the question: when comparing areas with high EP and high SE to areas with high EP but low SE, is the difference positively moderated by walkability? The analysis will use Allegheny County census data collected by US Census Bureau, EPA, and WPRDC. The EP variables were selected using WPRDC's recent Market Value Analysis (MVA) 2021, which has been shown to be positively associated with economic success [6]. The SE variables were selected using David Waddington's research at Census Bureau Statistics Measure Equity Gaps Across Demographic Groups [2].

METHEDOLOGY

1. Download census tract shapefiles for Allegheny County
2. Determine variables that describe the economic performance (EP), social equity (SE) and walkability (W)
3. Find and download identified EP and SE variables by census tract

Variable	Source	Description
EP_1	DP03	Median Household Income 2020
EP_2	MVA	Median Sale Price 2017-2019
EP_3	MVA	Sale Price Variance 2017-2019
EP_4	MVA	Mortgage Foreclosures 2017-2019, as % Owner Occupied Housing Units
EP_5	MVA	% Residential Parcels Built 2016-2019
EP_6	MVA	% Homes in Very Poor Condition
EP_7	MVA	% Parcels with Housing Inspection Violations
EP_8	MVA	% Owner Occupied Housing Units 2015-2019
EP_9	MVA	% Rental Units Receiving Subsidy
EP_{10}	MVA	% Residential Area Part of Vacant Lot
SE_1	S1501	% Population with Bachelor's Degree by Race
SE_2	S1501	Median Income Pay Gap by Gender
SE_3	S1701	% Population Below Poverty by Race
SE_4	B02001	% Change in Race 2010 to 2020
SE_5	B19083	Gini Index
SE_6	B25071	Gross Rent as % of Income by Race

4. Transform each SE variables from the two values into a single composite value. Remember, each variable above will have two different data values. For example, median income by gender will have a value for female and male. To determine either high or low SE, there is a need to turn those two values into one value allowing us to say high = good, low = bad. Therefore, create a composite variable with score between 0 and 100, excluding Gini index as it's already a composite variable and Gross rent as it's already a number between 0 and 100.
    a. $Z = ABS(X - Y)$; X = either white, male, or 2020 | Y = either non-white, female, 2010
    b. Standardize $\gg \theta = \frac{Z - \mu_Z}{\sigma_Z}$
    c. Find the $\Pr(\theta)$ given the data is normally distributed $\mu = 0, \sigma = 1$
    d. Linearly adjust the range from values between 0.0 and 0.50 to a range between 0 and 100.
    $NewValue = (((OldValue - OldMin) * (NewMax - NewMin)) / (OldMax - OldMin)) + NewMin$
5. Create an individual census tract GIS layer for EP and SE
6. Join EP and SE variables with their respective layers by census tract
7. Download the National Walkability Index shapefile for the W variable
8. Dissolve the National Walkability Index to census tract, data arrives by census block group

Field Calculations
Sum Total land area
Sum Population
Sum Housing units
Sum Households
Sum Count of workers
Mean 8-tier employment entropy
Mean Employment and household entropy
Mean intersection density
Mean Distance from population-weighted centroid to nearest transit stop
Mean Walkability index
Median Home Sale Price

9. Run kmeans clustering on EP variables and describe the clusters (5 clusters with 1 being Not Applicable)
10. Run kmeans clustering on SE metrics and describe the clusters (4 clusters)
11. Calculated summary statistics by EP and SE Cluster ID, using the mean of each variable to understand cluster centroids
12. Determine rankings of EP and SE (Platinum, Gold, Silver, Bronze)

13. Visualize each census tract by EP and SE rankings, produce map layouts
14. Visualize each census tract by W, produce map layouts
15. Export data into Excel, standardize EP and SE variables. Higher values = good, lower values = bad. For variables where the opposite is true, multiple their standardized values by -1.
16. Derive a single EP and SE value based on the average of all their respective variables
17. Create a summary file including EP, SE, and W single value by census tract
18. Import into RStudio, analyze the relationships using a linear regression model for p-value and R-squared

$$lm(EP \sim SE, data = df)$$

$$lm(EP \sim W, data = df)$$

$$lm(EP \sim SE * W, data = df)$$

$$lm(EP \sim SE + W, data = df)$$

$$lm(SE \sim W, data = df)$$

RESULTS:

Economic Performance

The section will start with describing the kmeans cluster analysis. The economic performance clusters are displayed below. The cluster used 10 variables: median household income, median sale price, variance sale price 2017 to 2019, % owner occupied homes, foreclosures, % poor condition homes, % homes with violations after inspection, new parcels created, vacancy rates, and subsidized homes. Figure 1 shows the cluster results by census tract, Table 1 shows mean values for each cluster by variable. While Table 1 shows the raw values determined by kmeans clustering, Table 2 shows the standardized data with inverse relationships on values where low = good vs the typical high = good. Table 2 best describes how the rankings by cluster were determined. As shown in Table 2, the group marked as Platinum lead the way in almost all variables: highest median income, highest sale price, least sale variance, 2nd highest owner-occupied housing, most new parcels, least violations, least homes in poor condition, and least vacant lots. The group marked as Gold took 2nd on all levels except owner occupied housing where is scored the highest. Silver and Bronze groups were similar, with the biggest differences around variance in sale price, % of subsidized homes, and homes with inspection violations.

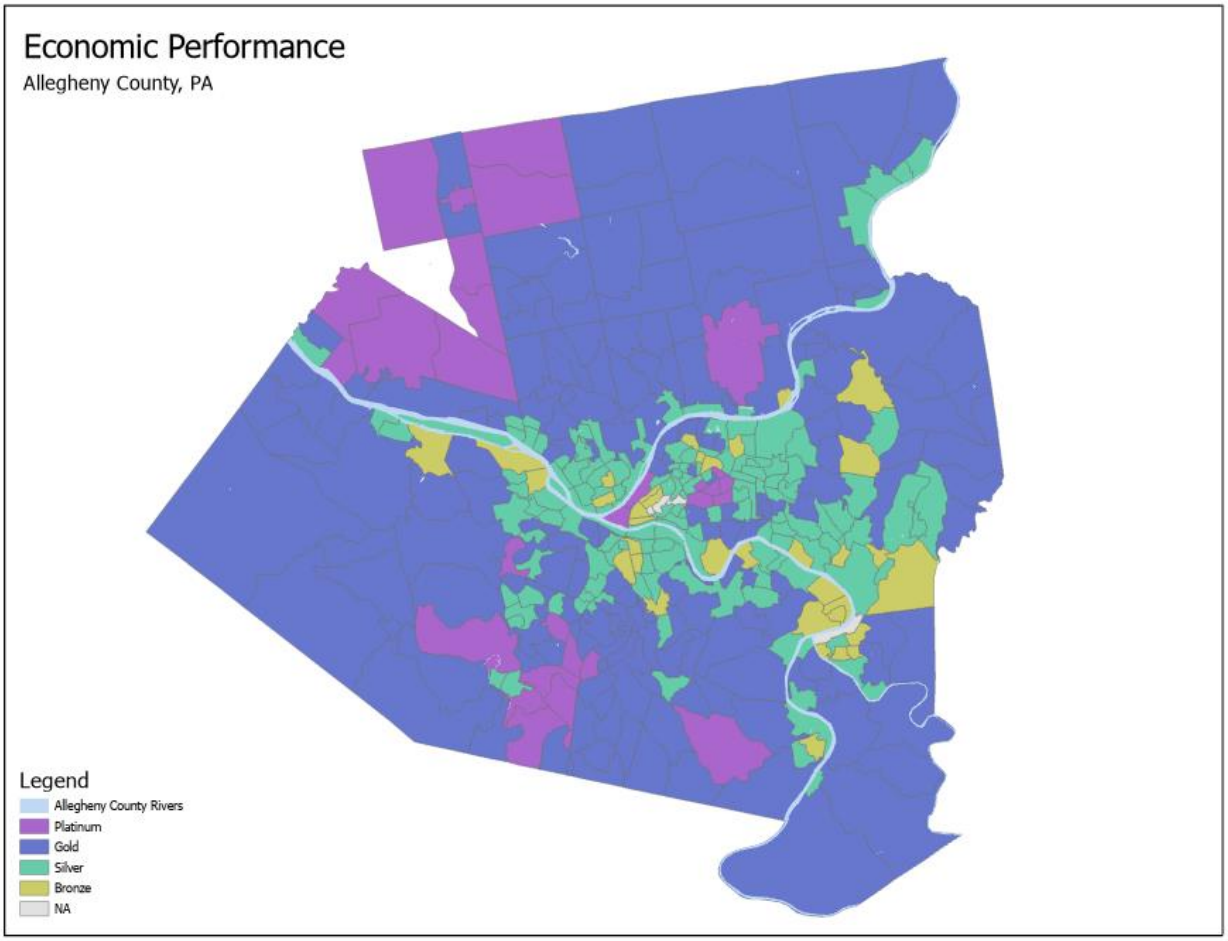

Figure 1: Economic Performance kmeans Clustering

CLUSTER ID	FREQUENCY	MEAN INCOME MEDIAN	MEAN SALE	MEAN SALE VAR	MEAN OWNER OCC	MEAN NEW PARCELS	MEAN SUBSIDIZED	MEAN VIOLATION	MEAN FORECLOSE	MEAN CONDITION
GOLD	180	76,161	175,975	-8.778	0.772	0.008	0.030	0.003	0.012	0.004
PLATINUM	29	136,345	413,617	-143.176	0.757	-86.175	0.026	0.001	-86.201	-143.677
NA	4	11,460	625	-1249.874	0.293	0.254	0.416	0.211	0.000	0.028
SILVER	145	39,852	80,620	0.509	0.399	0.003	0.083	0.010	0.016	0.011
BRONZE	35	36,764	56,257	-444.824	0.459	0.002	0.453	0.019	-255.083	-142.821

Table 1: Economic Performance kmean clustering, average value across variables

CLUSTER ID	FREQ	STAND_INC MEDIAN	STAND MSP1719	STAND VSP1719	STAND PHHOO	STAND PNROFRCNT	STAND PROSUBHH	STAND PVIOLADDRE	STAND PFORC1719	STAND PCOND_FLAG	STAND PVACLOT
PLATINUM	29	2.171	2.270	-0.119	0.586	0.408	0.408	0.342	0.655	0.406	-0.313
GOLD	180	0.389	0.255	-0.093	0.640	-0.034	0.384	0.243	0.141	0.261	0.197
SILVER	145	-0.687	-0.553	0.025	-0.753	-0.130	0.051	-0.027	-0.236	-0.058	-0.170
BRONZE	35	-0.778	-0.760	0.699	-0.529	-0.136	-2.288	-0.443	-0.419	-1.333	-0.006
NA	4	-1.528	-1.232	-1.938	-1.151	4.486	-2.058	-8.548	1.137	-0.913	-0.360

Table 2: Economic Performance kmean clustering, standardized then averaged value across variables

Social Equity

The social equity clusters are displayed below. The cluster used 6 variables: % population holding bachelor's degree by race, median income pay gap by gender, % population below poverty by race, % change in race 2010 to 2020, Gini index, and gross rent as percentage of income by race. Figure 2 shows the cluster results by census tract, Table 3 shows mean values for each cluster by variable. Remember, all of these variables are a range between 0 and 100 (except Gini index and Gross Rent as % of Income) where 100 is optimal SE between either race or gender. While Table 3 shows the raw values determined by kmeans clustering, Table 4 shows the standardized data with inverse relationships on values where low = good vs the typical high = good. While EP focused on a detailed interpretation of numerical values, SE discussions need to focus on the combination of the variables used. It is more important to note the overarching themes that kmeans clustering determined vs solely numerical interpretations, the clustering is too nuanced. After analyzing the clusters and variables, a few themes presented themselves. The Bronze group was clearly the areas within Allegheny County that were going through gentrification. The Bronze cluster had the highest displacement rates for people of color between 2010 and 2020, the largest Gini index value, and the worst rates for people of color being the only individuals below the poverty line. The Bronze group also had the 2nd lowest rate for education attainment equality. Based on the combination of the above results, it's reasonable to infer that these areas are heavily affected by gentrification. The Platinum group was determined to be Allegheny County's working-class individuals. This inference was mainly determined by the race equality between people below the poverty line, the lowest race change in 10 years, and the relative equality of gender pay with education attainment. The Silver cluster is likely suburban, white families due to the large disparity with people of color and highest rate of gender pay gap (men being the main wage earners). This is also inferred due to the low race change and low rate of rent to income. Finally, the Gold cluster is likely the old neighborhoods comprised of middle to upper class individuals. The inference is made due to the highest equality in education attainment, highest gender pay equality, and the lowest rate of rent to income. The analysis might be finding these relationships due to the high proportion of professors, nurses, and doctors in those census tracts, where occupations tend to pay well but also draw diverse groups of people.

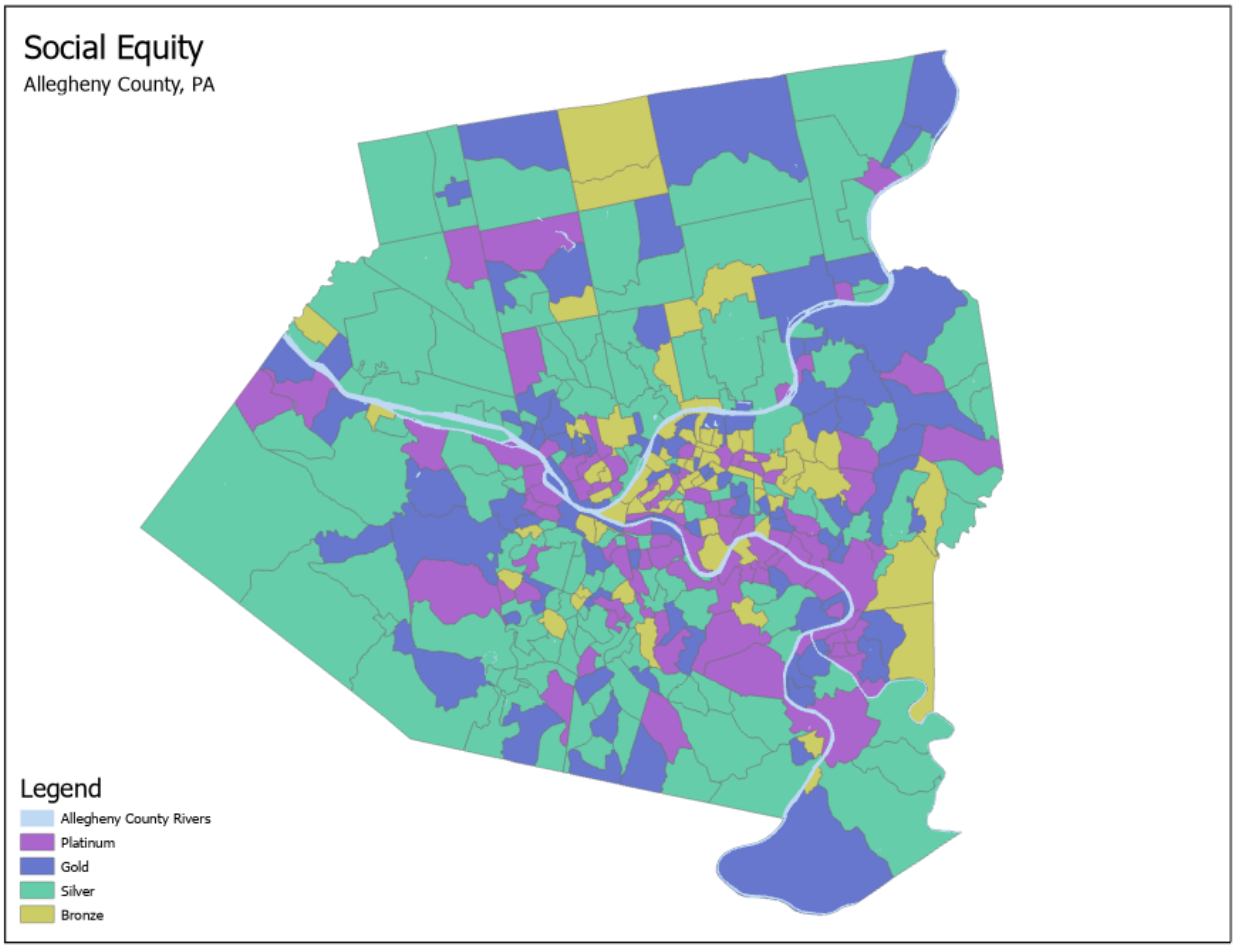

Figure 2: Social Equity kmeans Clustering

CLUSTER ID	FREQUENCY	MEAN EDUCATION ATTAINMENT INDEX	MEAN GENDER PAY GAP INDEX	MEAN POVERTY INDEX	MEAN RACE CHANGE	MEAN GINI INDEX	MEAN RENT BY INCOME
SILVER	139	16.122408	32.922637	10.086331	63.208633	0.400658	23.453237
PLATINUM	91	47.371796	46.245672	37.274725	72.714286	0.459271	33.945055
GOLD	94	80.151058	54.776611	9.042553	57.468085	0.345852	21.659574
BRONZE	70	22.661417	47.48263	18.628571	35.3	0.474043	32.585714

Table 3: Social Equity kmean clustering, average value across variables

CLUSTER ID	FREQ	STAND EDUC INDEX	STAND GENDER PAY INDEX	STAND POVERTY INDEX	STAND RACE CHANGE INDEX	STAND GINI INDEX	STAND RENT BY INCOME
PLATINUM	91	0.231	0.075	1.035	0.692	-0.446	-0.679
GOLD	94	1.227	0.337	-0.453	-0.082	0.675	0.535
SILVER	139	-0.719	-0.334	-0.398	0.210	0.133	0.357
BRONZE	70	-0.520	0.113	0.052	-1.207	-0.592	-0.545

Table 4: Social Equity kmean clustering, standardized then averaged value across variables

Walkability

The walkability analysis was determined through quantile analysis on the EPA's National Walkability Index (NWI) provided value, which was averaged by the census block groups per census tract. The NWI variable is determined through analyzing the variables: count of housing units, count of workers, variation in employment types (retail, office, industrial) termed employment entropy, employment and household entropy, intersection density, and distance from population-weighted centroid to nearest transit stop. While the majority of Pittsburgh city is walkable, we are able to identify a few key areas that are surprising unwalkable. This is likely due to the lack of employment entropy in these areas, however the hill landscape could also play a large role. Finally, we can see the worst NWI values arise in the suburban areas on Pittsburgh.

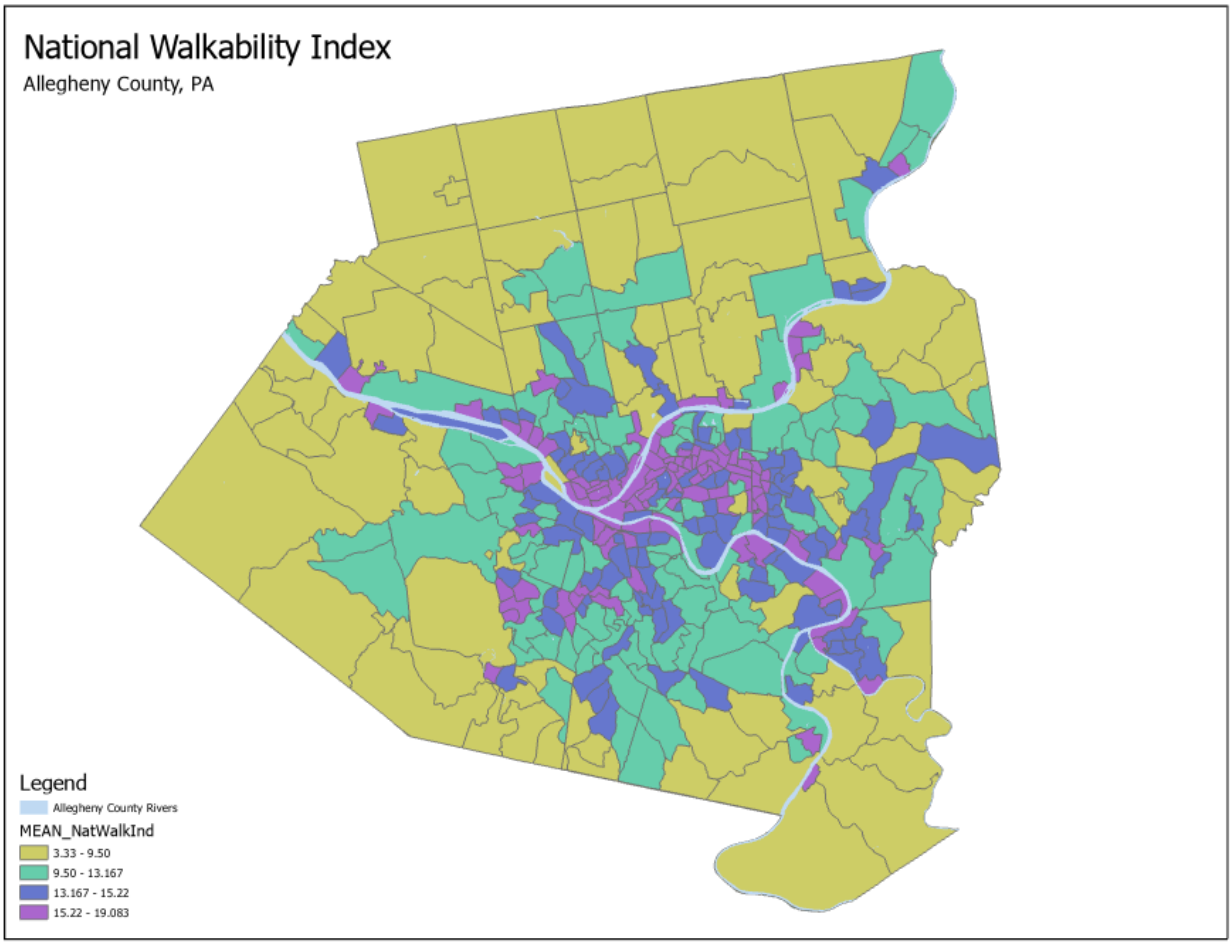

Figure 3: National Walkability Index, Quantile division with 4 classes

Composite Value Analyses

Finally, the results discussion will focus on analyzing the interactions between these key indicators: economic performance, social equity, and walkability. In order to simplify the interaction analysis, a single composite variable for both economic performance and social equity was calculated. The composite value was calculated using the standardized values for each variable, then taking the average across all standardized values. For example, each of the six variables were standardized for SE. Then, for items where the reverse interaction is preferred (i.e., when low = good instead of high = good), the standardized value was multiplied by -1. Finally, the average score was calculated across all six variables. That composite

value was used for the SE score. The same process was completed for EP, an average-standardized value across all 10 variables. The two composite variables for EP and SE, along with the NWI variable, by census block were then imported into RStudio for analyses.

Figure 4 and Figure 5 show how the composite values calculated for EP and SE align with their respective ranks from clustering, meaning the calculation transformation validated the kmeans clustering method results and interpretation.

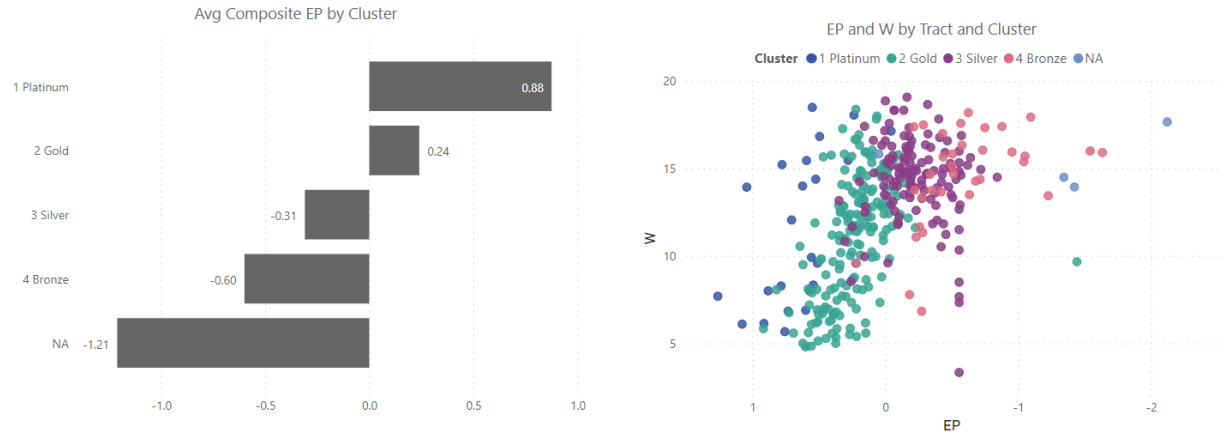

Figure 4: EP Composite Value by Cluster

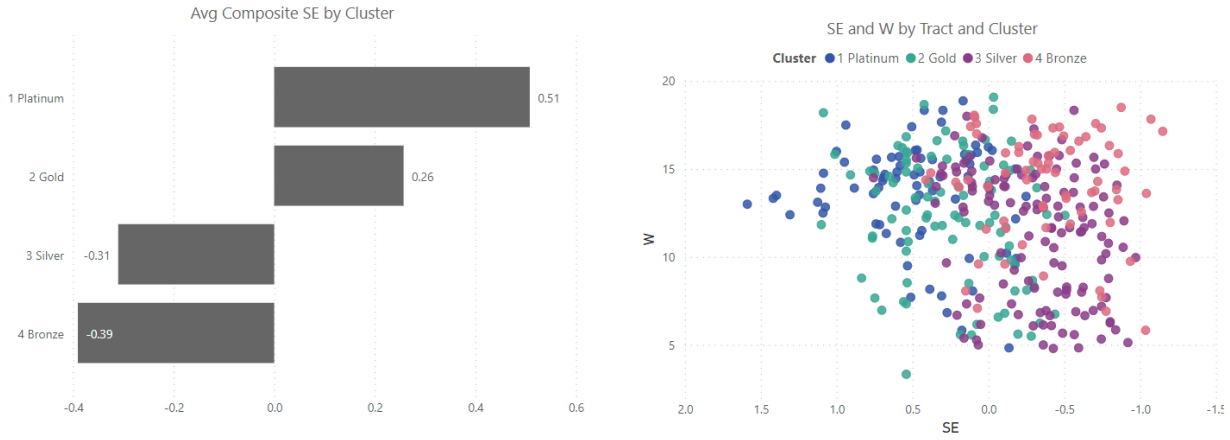

Figure 5: SE Composite Value by Cluster

Figure 6 shows that SE is negatively correlated with EP, p-value = <2e-16 and R-squared = 0.1905. The relationships between the two is fairly large, with one unit increase in SE, EP goes down -0.44. Considering the entire range of EP is between [-1.5, 1.5], the results are quite unfortunate. For census tracts that have high SE, they will have low SE. Referring to Figure 7, similar findings were found between W being negatively correlated with EP, p-value = <2e-16 and R-squared = 0.2515. The key outcome this project was hoping to prove is in Figure 8, does W moderate the interactions between EP and SE? The results found that the p-value = 0.5487, which means the results are not significant. The conclusion should be that W will not moderate the negative correlation between EP and SE. The final two figures go to further understanding our key indicators. Figure 9 shows the model that uses both W and SE to explain EP, the results found that W and SE continue to be significant, p-value = <2e-16 and p-value = 1.19e-14, respectively

and R-squared 0.3632. These results show that as EP increases, W and SE decrease independently. These findings motivated the final model, to see if there is a positive correlation between W and SE. The results found that W is positively correlated with SE p-value = 0.000218 and R-squared = 0.03425. While the R-squared isn't large, the findings were significant meaning areas in Pittsburgh that are more walkable with have higher SE. That could mean that creating policies for increasing walkability might in-turn increase the social equity of census tracts.

```
Call:
lm(formula = EP ~ SE, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.97847 -0.22411  0.02417  0.24110  3.15660

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.004741   0.024484  -0.194    0.847
SE          -0.448563   0.046458  -9.655   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4854 on 391 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.1925,    Adjusted R-squared:  0.1905
F-statistic: 93.23 on 1 and 391 DF,  p-value: < 2.2e-16
```

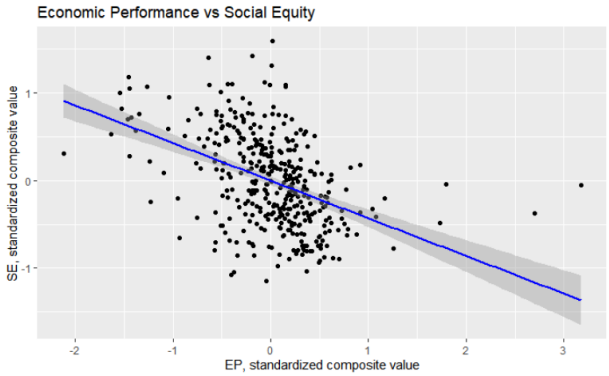

Figure 6: Linear regression EP and SE, p-value = <2ep16 | R-squared = 0.1905

```
Call:
lm(formula = EP ~ W, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.81260 -0.15462  0.03967  0.22640  1.11904

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.790648   0.073109   10.81   <2e-16 ***
W           -0.062263   0.005599  -11.12   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3794 on 364 degrees of freedom
  (28 observations deleted due to missingness)
Multiple R-squared:  0.2536,    Adjusted R-squared:  0.2515
F-statistic: 123.6 on 1 and 364 DF,  p-value: < 2.2e-16
```

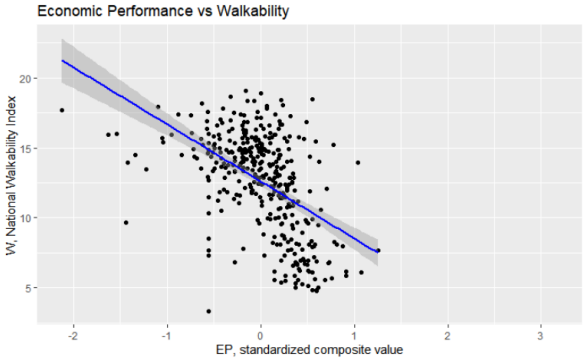

Figure 7: Linear regression EP and W, p-value = <2ep16 | R-squared = 0.2515

```
Call:
lm(formula = EP ~ SE * W, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.77006 -0.15241  0.04045  0.20873  1.02568

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.667966   0.072521   9.211   <2e-16 ***
SE          -0.375379   0.147998  -2.536   0.0116 *
W           -0.053227   0.005476  -9.721   <2e-16 ***
SE:W         0.006699   0.011161   0.600   0.5487
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3503 on 362 degrees of freedom
  (28 observations deleted due to missingness)
Multiple R-squared:  0.3673,    Adjusted R-squared:  0.362
F-statistic: 70.04 on 3 and 362 DF,  p-value: < 2.2e-16
```

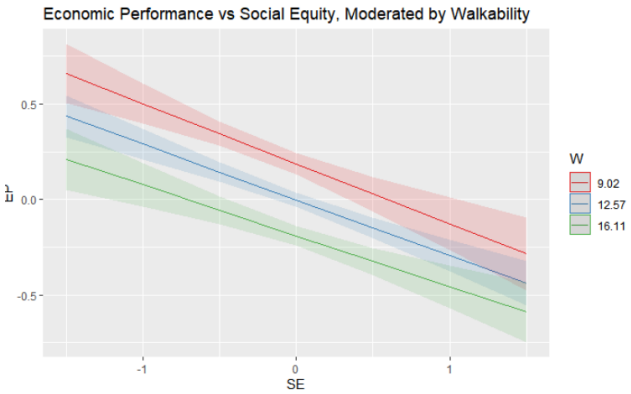

*Figure 8: Linear regression EP and SE*W, p-value = 0.5487 | R-squared = 0.362*

```
Call:
lm(formula = EP ~ SE + W, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.75794 -0.15463  0.04048  0.20617  1.02220

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.681653   0.068782    9.91  < 2e-16 ***
SE          -0.289208   0.035924   -8.05 1.19e-14 ***
W           -0.054125   0.005263  -10.28  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.35 on 363 degrees of freedom
  (28 observations deleted due to missingness)
Multiple R-squared:  0.3666,    Adjusted R-squared:  0.3632
F-statistic: 105.1 on 2 and 363 DF,  p-value: < 2.2e-16
```

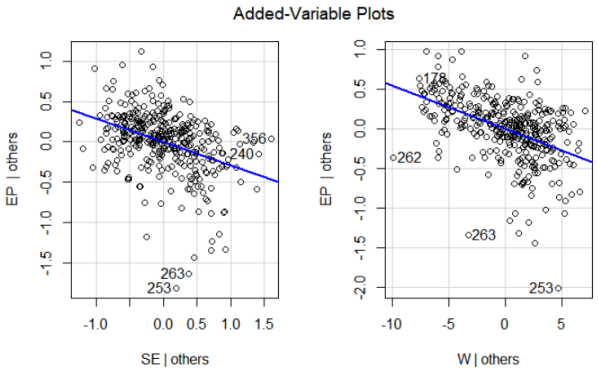

Figure 9: Linear regression EP and SE, p-value = 1.19e-14| EP and W, p-value <2e-16 | R-squared = 0.3632

```
Call:
lm(formula = SE ~ W, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.25527 -0.37548 -0.02533  0.37170  1.60000

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.376875   0.098391  -3.830 0.000151 ***
W            0.028140   0.007536   3.734 0.000218 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5106 on 364 degrees of freedom
  (28 observations deleted due to missingness)
Multiple R-squared:  0.0369,    Adjusted R-squared:  0.03425
F-statistic: 13.94 on 1 and 364 DF,  p-value: 0.0002185
```

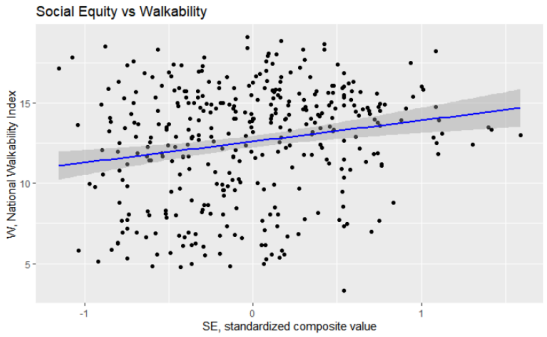

Figure 10: Linear regression SE and W, p-value = 0.000218 | R-squared = 0.03425

CONCLUSION AND FUTURE WORK

While the project didn't prove that higher W explains why tracts have both high EP and high SE, the project was able to find a statistically significant, positive correlation between W and SE. The positive relationship between W and SE justify the need for further analysis by policy makers. These results could justify the improvement of W in targeted census tracts, expecting to drive an increase in SE. Additionally, the analysis shows that a free market within Pittsburgh doesn't drive an increase in both EP and SE. If anything, the results seem to imply the opposite. Also, it is worth noting that high EP was mainly related to the census tracts outside of Pittsburgh. There seems to be a need for policy to drive the increase in EP in Pittsburgh.

REFERENCES

[1] Leinberger, C. B. (2012). *DC: The WalkUP Wake-Up Call.* George Washington University, School of Business.

[2] Waddington, David. "Census Bureau Statistics Measure Equity Gaps across Demographic Groups." *Census.gov*, 13 Apr. 2022, https://www.census.gov/library/stories/2021/09/understanding-equity-through-census-bureau-data.html.

[3] U.S. Department of Transportation, Bureau of Transportation Statistics. (n.d.), *Transportation Economic Trends*, available at www.bts.gov/product/transportation-economic-trends.

[4] Bureau, US Census. “Tiger/Line Shapefiles.” *Census.gov*, US Census Bureau, 16 Dec. 2021, https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.2020.html.

[5] Explore Census Data. https://data.census.gov/cedsci/. Accessed 29 Apr. 2022.

[6] Goldstein, Ira. *Allegheny County and City of Pittsburgh Market Value Analysis (MVA) 2021*.

[7] “Walkability Index.” *EPA*, Environmental Protection Agency, 13 May 2021, https://edg.epa.gov/metadata/catalog/search/resource/details.page?uuid=%7B251AFDD9-23A7-4068-9B27-A3048A7E6012%7D.

DATASETS:

EPA

https://www.epa.gov/smartgrowth/national-walkability-index-user-guide-and-methodology#:~:text=The%20National%20Walkability%20Index%20(2021,to%20rank%20the%20block%20groups.

- NATIONAL WALKABILITY INDEX
- Publication Date: 05/13/2021

US Census Bureau, filtered 2020 and All Census Tracts within Allegheny County

http://data.census.gov

- DP03 SELECTED ECONOMIC CHARACTERISTICS
- S1501 EDUCATION ATTAINMENT
- B25071 MEDIAN GROSS RENT AS A PERCENTAGE OF HOUSEHOLD INCOME IN THE PAST 12 MONTHS (DOLLARS)
- S1701 POVERTY STATUS IN THE PAST 12 MONTHS
- B19013 MEDIAN HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2020 INFLATION-ADJUSTED DOLLARS)
- S1903 MEDIAN INCOME IN THE PAST 12 MONTHS (IN 2020 INFLATION-ADJUSTED DOLLARS)
- B02001 RACE

US Census Bureau, filtered 2010 and All Census Tracts within Allegheny County

http://data.census.gov

- B02001 RACE

US Census Bureau, TIGER/Line Shapefiles, filtered 2020

https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.2020.html

- Pennsylvania, Allegheny County

WPRDC

https://data.wprdc.org/dataset/market-value-analysis-2021

- Pittsburgh Market Value Analysis 2021