PHILADELPHIA WORKS INC.
Building a Skilled and Thriving Workforce

How Does Distance Impact Enrollment?

Using the sf() Package in R to Investigate Enrollment Rates by Distance from CareerLink Centers

May 2023 | Abigail Jones and Sam Fraley

The Research & Data Team at Philadelphia Works is continually providing support across the organization for a broad scope of data and analytical topics. Among other focuses, R&D aims to provide data-driven insights to help Philadelphia Works better serve our customer base in Philadelphia. As a special topic in our research agenda, we recently conducted & presented analysis on how distance impacts our customers' enrollment rates.

Enrollment rates are a key metric in gauging the ability of centers to engage and retain customers in our programs, and staff is continually investigating ways to increase customer enrollment & retention to better improve outcomes. However, to better serve customers and improve enrollment rates, we must investigate what the underlying causes of unenrollment or disconnection are from our customers to CareerLink centers. One question that has come up, and that inspired this research, is **how does distance to the nearest CareerLink center impact a customer's likelihood of being enrolled?**

After conducting our analysis, we came away with three key findings:

1. **Distance alone may not predict the likelihood of a customer to be enrolled.** Enrollment rates vary greatly across zip codes, and there is not a clear enough correlation to conclude that increasing distance lowers enrollment rates.

2. **Distance may be related to other factors that influence enrollment for a customer.** Despite lack of correlation, we do see some concentration of low enrollment rate zip codes further from centers. Customers that rely on public transit may be less likely to enroll if centers are not easily accessible by transit, which could be related to distance.

3. **Further research can be done to better address the question.** We suggest looking specifically at workers who became disconnected from the system, to better see if there is a common spatial factor influencing their unenrollment or disconnection.

Methodology: Distance & Nearest Center

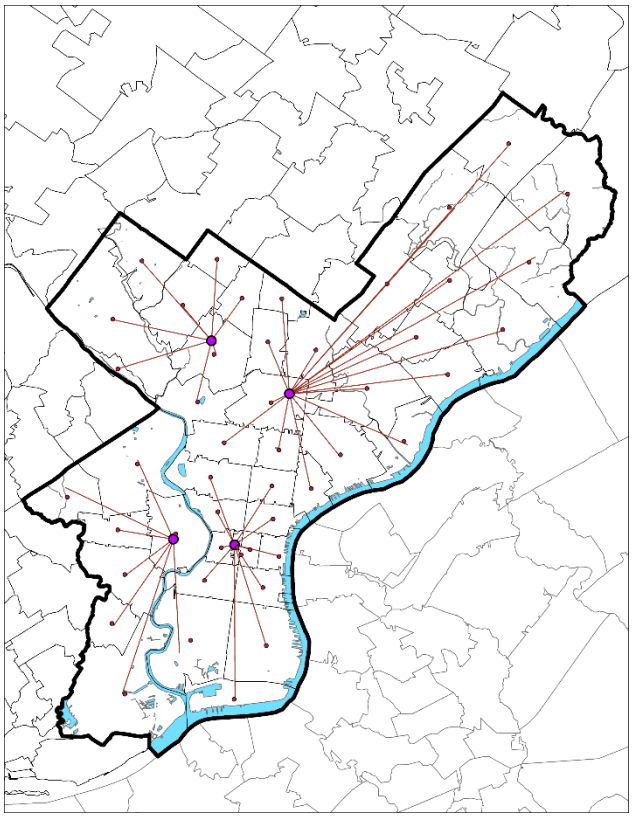
To conduct our analysis, we utilized customer data from the first half of program year 2022. To answer our broad research question, we first had to determine a few things from the customer data:

1. **What is the enrollment rate by zip code?** Using the zip code field in customer data, we aggregate up to determine the total numbers of customers, those that enrolled, and thus calculate the total enrollment.
2. **What is the nearest center to a customer?** While we do not have the exact home location of each customer, we approximated the nearest center by looking at the distance of all four centers to the "centroid"- or geographic center point - of the customer's home zip code. We then selected the center with the smallest distance and identified this as the "near" center.
3. **How far away is the customer?** Again, lack of specificity in customer location led us to approximating the customer's distance to the nearest center. We calculated the distance from the centroid of the zip code the nearest center and identified this as the distance to be used in analysis.

While using the centroid method to identify the nearest center and calculate distance is a generalized model that does have some short comings, we decided it was the most efficient way to conduct this exploratory analysis that, hopefully, will inform a more targeted and specified analysis in the future.

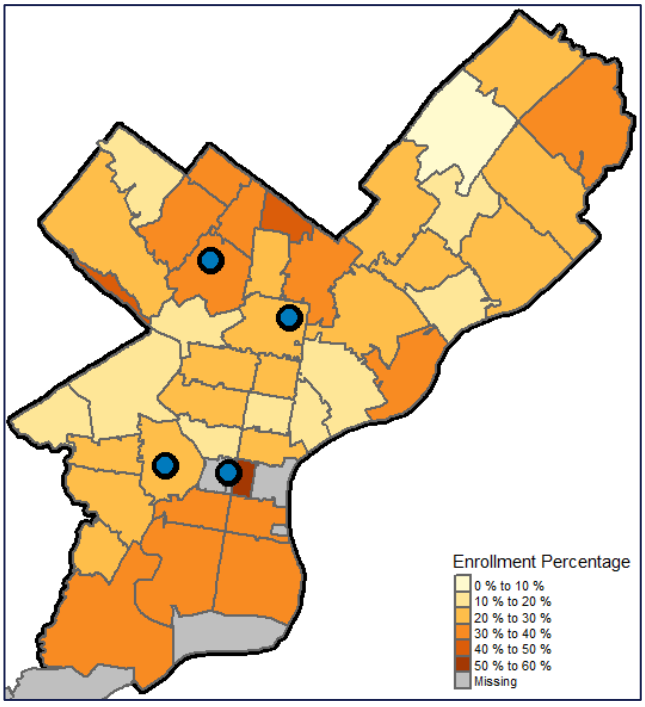
Finding the Nearest Center

Figure 1: Zip Code Centroids & Linear Distance to Nearest CareerLink Center



Calculating Enrollment Rates

Figure 2: Enrollment Rate by Zip Code



Results

After calculating the enrollment rates and identifying the distance to the nearest center for each zip code, we first plotted the miles from nearest center against the enrollment percentage, and used the geom_smooth() function in the ggplot2 package to calculate a best fit line[1]. When we assess the plot, we see a wide range of error and non-linear relationship that makes it hard to draw a concrete conclusion about the relationship.

To test the relationship quantitatively, we employ a simple linear regression model using the lm() function in R. The resulting regression output is displayed in Table 1. The result of this simple regression gives us a negative coefficient on distance as a model for enrollment rates, but the **p-value of 0.101 tells us it is not significant at the 95% confidence level.** In addition, the multiple R squared value is very low, illustrating that distance alone **has weak explanatory power for enrollment rates.** In short, the regression output tells us using distance alone to predict enrollment rates **is not a very good model.**

Figure 3: Distance vs. Enrollment Plot

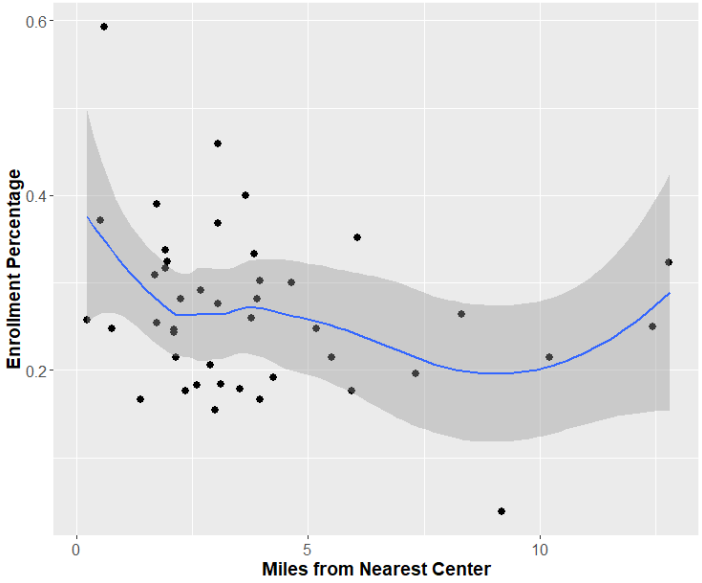


Table 1: Regression Output

Coefficient	Estimate	Std. Error	t. value	Pr(>|t|)
Intercept	3.00e-02	2.35e-03	12.74	7.7e-16*
Distance	-5.1e-08	3.01e-07	-1.68	0.101

Multiple R-Squared: 0.0644

1, The geom_smooth() package utilizes a “LOESS” or Local Regression method to create a best fit line. We use it to generally model the data.

While the regression model leaves much to be desired, we decide to conduct a high-level analysis to better understand what is happening on the extreme ends of the distance-enrollment rate plot. We decided to group the zip codes that are closest and furthest away from centers and investigate how their enrollment rates vary from the mean of the data set (Table 1). While this does not give us any ability to make predications or "model" distance and enrollment, it serves as explanatory analysis to be understand what is happening in our data.

Table 2: Enrollment Rates for Zip Codes

Zip Code	Distance	Enrollment Rate
19154	12.80 miles	32%
19116	12.43 miles	25%
19114	10.21 miles	21%
19115	9.18 miles	4%
Mean	**3.77 miles**	**27%**
19104	0.22 miles	26%
19144	0.52 miles	37%
19107	0.60 miles	59%
19140	0.76 miles	25%

When looking at Table 2, we see that 3 of the 4 closest zip codes have enrollment rates above the mean of 27%, and 4 out of 5 of the furthest zip codes have enrollment rates below this mean. While this may not give us a definite answer towards our research question, it does suggest that there is at least some relationship between distance and enrollment rate.

Conclusions & Next Steps

After conducting our research, we find that using distance alone to model or predict enrollment rates is not sufficient. We do find that zip codes closest to centers appear to have higher enrollment rates, and zip codes furthest away tend to have lower, but it is hard say exactly how much distance impacts enrollment rates.

We recommend a more specified analysis in the future. Our spatial aggregation to zip codes and centroid method to calculate distance leaves room for error, so future research should be done that utilizes more specific spatial or location data. In addition, looking specifically at the population who unenrolls or experiences exits of any type may be useful to identify if there is some spatial factor common across this population. On a high level, we understand that accessibility depends on more than the linear distance from a customer to a center. For example, access to and frequency of public transportation might explain enrollment rates better than our measure of distance. Future research should be done to consider how customers are getting to centers, how the relationship between operating hours and transit schedules might impact likeliness to come in, and other factors that might help us explain the relationship between distance, accessibility, and enrollment.

Appendix

Table 2: Full Zip Code & Enrollment Rate Data Table

Zip Code	Name of Nearest Center	Total Customers	Total Enrolled	Percent Enrolled	Distance (miles)
19139	West Philadelphia	592	144	24%	2.1
19143	West Philadelphia	564	159	28%	2.24
19131	West Philadelphia	528	97	18%	3.12
19140	North Philadelphia	315	78	25%	0.76
19124	North Philadelphia	311	64	21%	2.9
19104	West Philadelphia	303	78	26%	0.22
19134	North Philadelphia	285	44	15%	2.98
19132	North Philadelphia	275	76	28%	3.05
19144	Northwest Philadelphia	248	92	37%	0.52
19151	West Philadelphia	219	42	19%	4.25
19133	North Philadelphia	219	47	21%	2.15
19121	Suburban Station	213	62	29%	2.68
19142	West Philadelphia	206	58	28%	3.9
19120	North Philadelphia	190	64	34%	1.92
19138	Northwest Philadelphia	176	57	32%	1.95
19141	North Philadelphia	142	35	25%	2.1
19111	North Philadelphia	126	27	21%	5.5
19145	West Philadelphia	126	42	33%	3.83
19122	Suburban Station	115	21	18%	2.6
19149	North Philadelphia	113	28	25%	5.18
19146	Suburban Station	100	39	39%	1.74
19135	North Philadelphia	85	15	18%	5.95

Zip Code	Name of Nearest Center	Total Customers	Total Enrolled	Percent Enrolled	Distance (miles)
19123	Suburban Station	63	16	25%	1.73
19126	Northwest Philadelphia	61	28	46%	3.06
19150	Northwest Philadelphia	57	21	37%	3.06
19119	Northwest Philadelphia	55	17	31%	1.7
19136	North Philadelphia	53	14	26%	8.32
19152	North Philadelphia	46	9	20%	7.33
19148	Suburban Station	43	13	30%	3.96
19147	Suburban Station	41	13	32%	1.92
19153	West Philadelphia	37	13	35%	6.07
19129	Northwest Philadelphia	34	6	18%	2.35
19154	North Philadelphia	34	11	32%	12.8
19114	North Philadelphia	28	6	21%	10.21
19125	Suburban Station	28	5	18%	3.54
19107	Suburban Station	27	16	59%	0.6
19128	Northwest Philadelphia	27	7	26%	3.77
19115	North Philadelphia	26	1	4%	9.18
19116	North Philadelphia	16	4	25%	12.43
19130	Suburban Station	12	2	17%	1.38
19137	North Philadelphia	10	3	30%	4.64
19118	Northwest Philadelphia	6	1	17%	3.96
19127	Northwest Philadelphia	5	2	40%	3.65
19102	Suburban Station	NA	NA	NA	0.14
19113	West Philadelphia	NA	NA	NA	8.87
19106	Suburban Station	NA	NA	NA	1.71

Zip Code	Name of Nearest Center	Total Customers	Total Enrolled	Percent Enrolled	Distance (miles)
19109	Suburban Station	NA	NA	NA	0.44
19103	Suburban Station	NA	NA	NA	0.5
19112	Suburban Station	NA	NA	NA	5.78